

Seth Stephens-Davidowitz

MINDENKI HAZUDIK

AZ VAGY, AMIRE KLIKKELSZ

FORDÍTOTTA SZIEBERTH ÁDÁM

ATHENAEUM

A fordítás alapjául szolgáló mű
Seth Stephens-Davidowitz: Everybody Lies

Copyright © 2017 by Seth Stephens-Davidowitz
All rights reserved.

Hungarian translation @ Szieberth Ádám, 2019

Minden jog fenntartva.

2

301
391



Somogyi-könyvtár, Szeged



S000678808

Anyunak és Apunak

Kiadta az Athenaeum Kiadó,
az 1795-ben alapított Magyar Könyvkiadók
és Könyvterjesztők Egyesülésének tagja.
1086 Budapest, Dankó u. 4-8.
Tel.: 1-235-5030
www.athenaeum.hu
www.facebook.com/athenaeumkiado
athenaeumkiado2012@gmail.com

ISBN 978 963 293 868 4

Felelős kiadó: Szabó Tibor Benjámin
Felelős szerkesztő: Besze Barbara
Szerkesztette: Filippov Gábor
Műszaki vezető: Drótos Szilvia
Borítóterv: Földi Andrea
Nyomdai előkészítés: Tóth Viktor

Előszó

A filozófusok már régóta törik a fejüket a „cerebroszkóp” nevű mitikus készüléken, amely képes lenne képernyőre vetíteni egy ember gondolatait; a társadalomtudósok pedig legalább ilyen régóta kutakodnak olyan eszközök után, amelyek képesek feltárni az emberi természet működését. Kísérleti pszichológusi pályám során különböző módszerek jöttek divatba, majd lettek ejtve, én pedig kipróbáltam mindet: az értékelőskálákat, a reakcióidő, illetve a pupilatágulás mérését, a funkcionális agyi képalkotást, sőt még az epilepsziásokba beépített elektródákat is – a páciensek szívesen múltatták az időt egy-egy nyelvi kísérlet alanyai-ként, amíg a következő rohamukat várták...

Mindazonáltal ezek közül egyik módszer sem enged korlátlan betekintést az elme működésébe. A problémát az a brutális „vagy-vagy” dilemma okozza, hogy egy emberi gondolat rendszerint komplex állítás: nem csupán annyit gondolunk, hogy „valami oroszokról szólt”, mint Woody Allen, miután gyorsolvasással átfutotta a *Háború és békét*. Csakhogy a tudós számára igen kemény dió egy teljes fényében ragyogó, keszekusza, többdimenziós állítás elemzése. Persze ha valaki kiönti a szívét, érzékeljük a „tudatfolyam” végtelen gazdagságát, de a monológ korántsem ideális adathalmaz, ha tesztelni akarunk egy hipotézist.

Ha viszont azt tartjuk szem előtt, hogy könnyen kvantifikálható adatokat kapjunk, például a különböző szavak elhangzása utáni reakcióidőt vagy a képek által kiváltott bőrreakciót mérjük, akkor készíthetünk ugyan statisztikát, de a megismerés komplex szövetét egyetlen számmá pépesítjük. Még a legszofisztikáltabb agyi képalkotó módszerek is csak azt árulják el, hogy miként „nyúlik el” egy gondolat a háromdimenziós térben, de azt már nem, hogy miből áll össze az a bizonyos gondolat.

Mintha nem lenne épp elég bajuk a kezelhetőség kontra adatgazdagság dilemmájával, az emberi természet kutatóinak a „kis számok törvénye” is komoly fejfájást okoz. Amos Tversky és Daniel Kahneman nevezte el így azt a tévhitet, amely szerint egy populáció jellemző vonásait hűen tükrözi bármely, mégoly csekély egyedszámú minta. A leginkább a számokra összpontosító tudósok intuíciója is csődöt mond, ha azt kell megítélniük, hány alanyra van szükség egy tanulmányhoz annak érdekében, hogy a véletlenszerű egyéni sajátosságoktól és esetlegességektől elrugaszkodva, minden amerikaira, pláne minden *Homo sapiens*re érvényes általánosításokat lehessen levonni. Még kétségesebb a siker, ha a mintavétel során kényelmi szempontok érvényesülnek, például sörpénzt ajánlunk fel a kurzusainkra járó másodéveseknek.

Ez a könyv az emberi elme tanulmányozásának mérően újszerű módjáról szól. Az internetes keresésekből, illetve más online megnyilatkozásokból kinyert, irdatlan méretű adathalmaz, a „Big Data” nem cerebroszkóp, de Seth Stephens-Davidowitz rávilágít, hogy eddig példátlan mélységű bepillantást enged az emberek pszichéjébe. Az emberek a billentyűzet előtti magányukban roppant fura dolgokat ismernek be – olykor (például randioldalakon

vagy ha szakmai tanácsokra keresnek rá) azért, mert ennek van következménye a való életben, máskor pedig éppen azért, mert *nincs* következmény: úgy könnyíthetnek a lelkükön, úgy vallhatják meg vágyaikat, félelmeiket, hogy azzal egyetlen valódi személyből sem váltanak ki döbbenetet, rémületet vagy még rosszabbat. De akár van következmény, akár nincs, az emberek ilyenkor nem csupán megnyomnak egy gombot vagy elfordítanak egy kart, hanem begépelik a sok-sok trillió lehetséges karaktersorozat egyikét, hogy körvonalazzák gondolataikat a maguk robbanásveszélyes, kombinatorikus, mérhetetlen valójában. Sőt mi több, ezeket a digitális nyomokat könnyen összesíthető és analizálható formában terítik. Az élet legkülönbözőbb területeiről jönnek, és közben észrevétlen kísérletek alanyai vá válhatnak, amelyek során mind az ingerek módosítása, mind a válaszok csoportosítása valós időben történik. Ők pedig önként és dalolva szolgáltatják ezt az irdatlan mennyiségű adatot.

A *Mindenki hazudik* nem csupán egy elmélet igazolása: Stephens-Davidowitz felfedezései újra meg újra a fejük tetejére állították a hazámmal, illetve az emberi nemmel kapcsolatos prekonceptióimat. Honnan jött Donald Trump váratlanul magas támogatottsága? Amikor Ann Landers 1976-ban azt kérdezte olvasóitól, nem bánták-e meg, hogy gyereket vállaltak, és sokkolta, hogy a többség megbánta, lehet, hogy az önkényesen összeválogatott, nem reprezentatív minta vezette félre? Az internet okolható a kései 2010-es évek kissé redundáns elnevezésű válságáért, a „szűrőbuborékért”? Mi váltja ki a gyűlölet-bűncselekményeket? Azért keresnek vicceket az emberek, hogy felvidítsák magukat? És bár szeretem azt gondolni, hogy engem már semmi sem tud sokkolni, bőven sokkolt az, amit az

internet az emberi szexualitásról felfed: beleértve azt a felfedezést, hogy meghatározott számú nő minden hónapban rákeres arra, hogy „dugás plüssállatokkal”. Ezt az adatot soha nem hozták volna felszínre sem a reakcióidőt vagy a pupillatágulást mérő kísérletek, sem pedig a funkcionális agyi képalkotás.

A *Mindenki hazudik*ot mindenki élvezni fogja. Stephens-Davidowitz lankadatlan kíváncsiságtól fűtött, megnyerően okos könyve új utat mutat a 21. századi társadalomtudománynak. Ennek az emberi rögeszmékre nyíló, végtelenül izgalmas ablaknak a kitérője után ugyan kinek van szüksége cerebroszkópára?

Steven Pinker, 2017

Bevezetés

Egy forradalom körvonalai

Holtbiztos, hogy veszteni fog – mondták.

A 2016-os republikánus előválasztás előtt a közvéleménykutatók arra jutottak, hogy Donald Trumpnak az égvilágon semmi esélye. Trump mégiscsak számos kisebbségi csoportot megsértett. A felmérések és azok értelmezői azt mondták nekünk, hogy az efféle durvaságot csak kevés amerikai tartja elfogadhatónak.

Akkoriban a szakértők nagy többsége úgy vélte, hogy Trump elveszíti az elnökválasztást. A szavazni készülők közül túl sokan állították azt, hogy elriasztja őket a modora és a nézetei.

Pedig valójában voltak előjelei annak, hogy Trump mind az előválasztást, mind az elnökválasztást megnyerheti. Bizony, voltak – az interneten.

Internetes adatszakiértő vagyok. Nap mint nap követem a weben járó-kelő emberek digitális lábnyomait. A lenyomott-leütött gombok, billentyűk segítségével próbálom megérteni, mit is akarunk valójában, mit is teszünk valójában, kik is vagyunk valójában. Hadd magyarázzam el, hogy vágtam neki ennek a szokatlan útnak.

A történet a 2008-as elnökválasztással (mintha már ezer éve lett volna!), illetve egy, a társadalomtudósok által rég-

óta vitatott kérdéssel kezdődik. A kérdés így szól: mennyire jelentős Amerikában a faji előítélet?

Barack Obama volt minden idők első, valamely nagy párt által indított afroamerikai elnökjelöltje. Győzött – meglehetősen könnyedén. És a közvélemény-kutatási eredmények arra utaltak, hogy a faji hovatartozás nem játszott szerepet abban, hogyan szavaztak az amerikaiak. A Gallup például számos felmérést végzett Obama első megválasztása előtt és után. Az eredmény? Az amerikai szavazókat nagyrészt hidegen hagyta, hogy Barack Obama fekete. Ném sokkal az elnökválasztás után a Berkeley Egyetem két híres professzora más, szintén felméréseken alapuló adatokba ásta bele magát, jóval szofisztikáltabb adatbányászati technikákat alkalmazva. Ők is hasonló következtetésre jutottak.

Így aztán Obama elnöksége alatt a média nagy részében, illetve tudományos körökben ez közkeletű vélekedéssé vált. Azok a források, amelyekre a médiamunkások és a társadalomtudósok már vagy nyolcvan éve támaszkodtak a világ megértésében, azt mondták nekünk, hogy az amerikaiak túlnyomó többségét nem érdekli, hogy Obama fekete, amikor azt latolgatják, hogy ő legyen-e az elnökük.

Úgy tűnt, ez az ország, amelyet oly sokáig szennyezett a rabszolgaság és a Jim Crow-törvények,* végre felhagyott azzal, hogy a bőrük színe alapján ítélje meg az embereket. Úgy tűnt, a rasszizmus az utolsókat rúgja Amerikában. Sőt, egyes megmondóemberek odáig mentek, hogy kijelentték: posztrasszista társadalomban élünk.

* Jim Crow-törvények: az amerikai polgárháborút követően számos déli államban életbe lépő, diszkriminatív jogszabályok, amelyek a formális jogegyenlőség ellenére, a szövetségi állam jóváhagyásával egészen az 1960-as évekig fenntartották a fehérek és afroamerikaiak szegregációját, többek között a közoktatásban, a tömegközlekedésben és a hivatalokban. (A szerk.)

2012-ben posztgraduális közgazdászhallgató voltam – elveszve az életben, kiéve a szakmában, abban a magabiztos, sőt beképzelt hitben, hogy meglehetősen jól értem a világ működését, hogy mi jár az emberek fejében, mi foglalkoztatja őket a 21. században. Ami pedig az előítéleteség problémáját illeti, pszichológiai és politikatudományi olvasmányaim alapján elhittem magammal, hogy a nyílt rasszizmus az amerikaiak csekély hányadára korlátozódik – akiknek többsége konzervatív, a „Deep South”-on, vagyis a legdélebbi amerikai államokban élő republikánus.

Aztán felfedeztem a Google Trendset.

A Google Trends, amelyet különösebb csinnadratta nélkül indítottak útjára 2009-ben, olyan eszköz, amelynek segítségével megtudhatjuk, milyen gyakorisággal kerestek rá az emberek egy-egy szóra, kifejezésre különböző helyeken és különböző időpontokban. Szórakoztató termékként reklámozták, amely arra való, hogy az ember megvitassa a barátaival, melyik celeb a legnépszerűbb, vagy mi jött hirtelen divatba. A legkorábbi verziókhoz egy vicces intelmet is mellékeltek, amely szerint „doktori disszertációt azért nem kellene írni” ezen adatok alapján: ami azonnal arra sarkallt, hogy ezek alapján írjam meg a disszertáciomat.*

* Adataim nagy részének forrása a Google Trends. Mivel azonban így csak a különböző keresések relatív gyakoriságát tudom összevetni, de egy-egy adott keresés abszolút száma nem derül ki belőle, az adatbázist általában kiegészítettem a Google AdWordszel, amely arról is tájékoztat, hogy pontosan milyen gyakran kerestek rá valamire. Az esetek többségében tovább tudtam élesíteni a képet a saját, Google Trends-alapú algoritmusom segítségével, amelynek leírása *Essays Using Google Data* (Értekezések Google-adatok felhasználásával) című disszertáciomban, valamint a *Journal of Public Economics*-ben publikált, *The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data* (Mibe kerül egy fekete jelöltnek a fajgyűlölet? Bizonyítás Google-keresési adatok alapján) című tanulmányomban olvasható. A disszertáció, a cikk linkje, valamint az e könyvben bemutatott, eredeti kutatásaimhoz felhasznált adatok és kódolás részletes magyarázata megtalálható a honlapomon: sethsd.com.

A Google-keresések adatai akkoriban még nem tűntek megfelelő, azaz „komoly” tudományos kutatás céljára használható információforrásnak. A felmérésekkel ellentétben a Google-keresések adatait nem az emberi psziché megértését segítő eszköznek szánták. A Google-t arra találták ki, hogy az emberek mindenfélét megtudjanak a világról, nem arra, hogy a kutatók tudjanak meg mindenfélét az emberekről. De mint kiderült, a nyomok, amelyeket a tudást hajhászva hagyunk magunk után a neten, elképesztő mértékben árulkodók.

Más szóval: az emberek információkeresése önmagában is információ. Mint kiderült, az, hogy mikor és hol keresnek rá tényadatokra, idézetekre, viccekre, helyekre, személyekre, tárgyakra, illetve mikor és hol kérnek segítséget, sokkal többet árul el arról, hogy mit gondolnak valójában, mire vágnak valójában, mitől félnek valójában és mit tesznek valójában, mint azt bárki hitte volna. Különösen igaz ez azt figyelembe véve, hogy az emberek időnként nem is annyira faggatják a Google-t, mint inkább a bizalmukba fogadják: „Gyűlölöm a főnökömet.” „Részeg vagyok.” „Apám megütött.”

Az a mindennapos aktus, hogy egy-egy szót vagy kifejezést bepötyögünk egy kompakt, fehér téglalapba, az igazság apró nyomát hagyja maga után, amely idővel, megsokszorozódva, milliós nagyságrendben feltárja az élet legvalódibb tényeit. Az első szó, amelyet beírtam a Google Trendsbe, az „Isten” volt. Megtudtam, hogy a legtöbb olyan Google-keresés, amelyben szerepelt az „Isten” szó, Alabama, Mississippi és Arkansas államokhoz, vagyis a Bibliaövezet* köthető. Megtudtam továbbá, hogy az ilyen

* Az Egyesült Államok déli, döntően konzervatív, főként neoprotestáns felekezetekhez tartozó területei. (A szerk.)

keresések vasárnaponként a leggyakoribbak. Mindez egyáltalán nem volt meglepő, azt viszont már érdekesnek találtam, hogy a keresési adatok ilyen tisztán kirajzolódó mintázatot tudnak feltárni. Ezután megpróbálkoztam a „Knicks” szóval, és kiderült, hogy arra New York Cityben keresnek rá a legtöbben. Itt sincs semmi látnivaló. Aztán beírtam a saját nevemet. „Sajnáljuk. Nincs elegendő mennyiségű keresés” – tájékoztatott a Google Trends. Máris megtudtam tehát, hogy a Google Trends csak akkor szolgáltat adatokat, ha nagy tömegek keresnek rá ugyanarra.

De a Google-keresések hatalma nem abban rejlik, hogy megtudhatjuk: Isten lent, a „mély Délen” népszerű, a New York Knicks kosárlabdacsapata New Yorkban népszerű, én pedig nem vagyok népszerű sehol. Ezt bármilyen felmérésből megtudhatnánk. A Google-adatok igazi erejét az adja, hogy az emberek olyasmiket közölnek a gigantikus keresőmotorral, amit senki másnak nem árulnának el.

Vegyük például a szexet (egy olyan témát, amelyet ebben a könyvben később sokkal részletesebben fogok vizsgálni). Ha a szexuális életünkkel kapcsolatban az igazságra vagyunk kíváncsiak, a felmérésekben nemigen bízhatunk. Elemeztem a chicagói National Opinion Research Center (Országos Közvélemény-kutató Központ) által végzett általános társadalmi felmérés, a General Social Survey adatait, amelyet az amerikaiak viselkedésére vonatkozó egyik legbefolyásosabb, legtekintélyesebb információforrásnak tartanak. A heteroszexuális szexre vonatkozó felmérés szerint a nők azt állítják, hogy évente 55 alkalommal szexelnek, és az esetek 16 százalékában használnak óvszert. Ez összesen évi 1,1 milliárd óvszert jelent. Csakhogy a heteroszexuális férfiak azt állítják, hogy ők 1,6 milliárd óvszert használnak el évente. Márpedig a két számnak magától értetődően egyeznie kellene. Akkor hát ki mond igazat, a férfiak vagy a nők?

Mint kiderült, senki. A Nielsen nevű globális információszolgáltató cég szerint, amely a fogyasztói viselkedést méri, az évente eladott óvszerek száma nem éri el a 600 milliót. Tehát mindenki hazudik: az egyetlen különbség, hogy ki mennyire.

Ami azt illeti, a hazugság széles körben elterjedt szokás. Azok a férfiak, akik soha nem voltak házasok, állításuk szerint átlagosan 29 óvszert használnak el évente. Ha ez igaz lenne, az több óvszert jelentene, mint ahányat az Egyesült Államokban a házasoknak és a szingliknek együttvéve eladnak. Valószínűleg a házasok is túloznak a szex mennyiségét illetően. A 65 év alatti nők férfiak például azt mondják a közvélemény-kutatóknak, hogy átlagosan hetente egyszer szexelnek. Mindössze egy százalékuk állítja, hogy már egy éve szexmentes életet él. A férjes asszonyok valamivel kevesebb együttlétről számolnak be, de nem sokkal kevesebbről.

A Google-keresések viszont jóval kevésbé élénk – és úgy vélem, jóval pontosabb – képet adnak a házasság alatti nemi életéről. A Google-on az a leggyakoribb panasz a házassággal kapcsolatban, hogy nincs szex. A „szexmentes házasság” keresőkifejezés három és félszer gyakrabban fordul elő, mint a „boldogtalan házasság”, és nyolcszor gyakrabban, mint a „szerelem nélküli házasság”. Még az össze nem házasodott párok is elég sűrűn panaszkodnak a szex hiányára. A Google-keresésekben a „szexmentes kapcsolat” kifejezés gyakoriságát csak a „bántalmazó kapcsolat”-ra való keresések gyakorisága múlja felül. (Hangsúlyoznom kell, hogy ezek az adatok mindig névtelenek. A Google természetesen nem közöl információt egyetlen konkrét személy kereséseiről sem.)

És a Google-keresések olyan képet festettek Amerikáról, amely meghökkentően elütött a közvélemény-kutatások által felvázolt posztrasszista utópiától. Máig emlékszem a

pillanatra, amikor először írtam be a Google Trendsbe azt, hogy „nigger”. Mondják bátran, hogy naiv vagyok – de abból kiindulva, mennyire mérgező ez a szó, alacsony keresési mennyiségre számítottam. Hű, hát tévedtem! A „nigger” – vagy ennek többes száma, a „niggerek” – nagyjából ugyanannyi keresésben jelent meg az Egyesült Államokban, mint a „migrén”, a „közgazdász” és a „Lakers” szavak. Eszembe jutott, hogy talán a rapszövegek torzították el az eredményt. Hát nem. A rapperek szinte mindig a „nigga” („niggák”) szót használják. Akkor hát mi ösztönözte az amerikaiakat arra, hogy rákeressenek a „nigger”-re? Sokszor az, hogy az afroamerikaiakon gúnyolódó vicceket kerestek. Ami azt illeti, a „nigger” szót tartalmazó keresések csaknem 20 százaléka tartalmazta a „viccek” szót is. Gyakori keresőkifejezés volt még többek között a „buta niggerek”, valamint a „gyűlölöm a niggereket”.

Évente több millió ilyen keresés volt. Otthon, a négy fal között rengeteg amerikai tudakozódik sokkolóan rasszista tartalmak iránt. És minél tovább kutakodtam, annál nyugtalanítóbb információkra bukkantam.

Obama első megválasztásának estéjén, amikor a kommentárok zöme az új elnök méltatásáról szólt, illetve arról, hogy a győzelme történelmi jelentőségű, durván minden századik Google-keresésben, amely tartalmazta az „Obama” szót, szerepelt a „KKK” vagy a „nigger(ek)” is. Ez talán nem tűnik túl nagy aránynak, de gondoljanak csak bele, hány ezer nem rasszista oka lehetett annak, hogy az ember rákeressen erre a fiatal outsiderre, akinek ilyen bájos családja van, és éppen átveszi a világ legnagyobb hatalommal járó munkakörét! A választás estéjén a Stormfront nevű fehér nacionalista weboldalra, amely meglepően népszerű az Egyesült Államokban, tízszer annyian kerestek rá, illetve iratkoztak fel, mint egy átlagos napon. Egyes államokban

többször kerestek rá arra, hogy „nigger elnök”, mint arra, hogy „első fekete elnök”.

Volt ott valami sötétség és gyűlölet, amely a hagyományos források előtt rejtve maradt, de az emberek Google-kereséseiben elég látványosan megjelent.

Az ilyen keresések nehezen békíthetők össze egy olyan társadalom képével, ahol a rasszizmus jelentéktelen tényező. 2012-ben Donald J. Trump elsősorban üzletemberként és valóságshow-házigazdaként élt a fejében. Ugyanúgy nem sejtettem, ahogy senki más, hogy négy évvel később komoly elnökjelölt lesz belőle. Ám a fenti, rusnya Google-kereséseket már igazán nem nehéz összebékíteni egy olyan jelölt sikerével, aki – a bevándorlók elleni támadásaival, indulataival és gyűlölködésével – gyakran játszott rá az emberek legrosszabb hajlamaira.

A Google-keresések azt is elárulták, hogy amit a rasszizmus területi eloszlásáról gondoltunk, többnyire téves volt. A felmérések, illetve a közmegegyezés főként Délen és főként a republikánusok körében helyezte el a modern rasszizmust. Csakhogy az ország azon területei között, ahol a legmagasabb volt a rasszista keresések aránya, Nyugat-Virginia, Dél-Louisiana és Mississippi állam mellett megjelent New York állam északi része, Nyugat-Pennsylvania, Kelet-Ohio, a michigani iparvidék, valamint Illinois vidékies része is. A Google-keresések adatai szerint a valódi választóvonal nem Észak és Dél, hanem Kelet és Nyugat közt húzódott. A Mississippitől nyugatra nemigen találkozik ilyesmivel az ember. És a rasszizmus nem korlátozódott a republikánusokra. Sőt, ami azt illeti, a rasszista tartalmú keresések aránya nem volt magasabb azokon a helyeken, ahol magas volt a republikánusok aránya, mint ott, ahol a demokraták voltak túlsúlyban. Más szóval: a Google-ke-

resések segítettek átrajzolni az Egyesült Államok rasszizmustérképét, és ez a térkép egészen másképp nézett ki, mint amilyennek képzeltük. Lehet, hogy a déli republikánusok nagyobb valószínűséggel vállalják fel a rasszizmusukat. De rengeteg északi demokrata is hasonlóan viszonyul a kérdéshez.

Ez a térkép négy évvel később jelentősen hozzájárult Trump politikai sikerének magyarázatához.

2012-ben arra használtam a Google-keresések alapján kifejlesztett rasszizmustérképet, hogy átértékeljem Obama faji hovatartozásának szerepét. Az adatok egyértelműek voltak: azokban az országrészekben, ahol magas volt a rasszista tartalmú keresések száma, Obama lényegesen rosszabbul szerepelt, mint négy évvel korábban John Kerry, a demokraták fehér elnökjelöltje. A különbségre semmilyen más tényező nem adott magyarázatot: sem az iskolázottság szintje, sem a korosztályi összetétel, sem a templomba járás, sem a fegyverbirtoklás. A rasszista keresések száma nem jelezte előre egyetlen más demokrata jelölt gyenge szereplését sem. Egyedül Obamáét.

És az eredmények óriási hatásra utaltak. Obama országosan durván 4 százalékpontot veszített a nyílt rasszizmus miatt. Ez jóval nagyobb arány, mint amit bármilyen felmérés alapján várhattunk volna. Barack Obamát persze így is elnökké választották, majd újraválasztották, amit néhány, a demokraták számára roppant kedvező körülmény is elősegített. De jócskán több hátrányt kellett leküzdenie, mint azt a hagyományos adatforrásokra támaszkodók – azaz nagyjából mindenki – gondolták volna. Épp elég rasszista volt ahhoz, hogy eldöntsenek egy előválasztást, vagy átbilentsék egy elnökválasztás eredményét egy olyan évben, amikor a körülmények nem kedveztek ennyire a demokratáknak.

A tanulmányomat elsőre öt tudományos folyóirat is elutasította. Ha elnéznek nekem egy kis morgolódást: cikkeim több szakmai bírálója azt mondta, egyszerűen nem hihető, hogy ilyen sok amerikai rejteget ennyire durva rasszizmust a lelkében. Ez egészen egyszerűen nem volt összeegyeztethető mindazzal, amit akkoriban az emberek előszeretettel mondogattak. Ráadásul a Google-keresések igencsak bizarr adathalmaznak tűntek...

Ma azonban, miután tanúi voltunk Donald J. Trump elnöki beiktatásának, az eredményeim már jóval hihetőbbnek tűnnek.

Minél tovább tanulmányoztam a Google-t, annál világosabbá vált, hogy rengeteg olyan információt tartalmaz, amelyekre a közvélemény-kutatások nem derítenek fényt, és amelyek segítségével – sok-sok más téma mellett – egy választást is könnyebben megérthetünk.

Információkat nyerhetünk például arról, hogy ténylegesen ki fog elmenni szavazni. A választás előtti napokban végzett felméréseken a nem szavazó állampolgárok több mint fele azt válaszolja, hogy szándékában áll voksolni, ami torzítja a részvételi arányra vonatkozó becsléseinket. Az viszont, hogy hetekkel a választások előtt hányan keresnek rá arra, hogy „hogyan kell szavazni”, illetve „hol kell szavazni”, képes pontosan előre jelezni, hogy az országnak mely részeiben várható magas részvételi arány.

Sőt, talán még arról is lehetnek ott információk, hogy ki-re fognak szavazni az emberek. De hát hogy jósolhatnánk meg pusztán a Google-kereséseik alapján, hogy melyik jelöltre adják majd voksukat? Nyilván nem elég, ha azt vizsgáljuk, hogy melyik jelöltre keresnek rá a leggyakrabban. Sokan azért keresnek rá valamelyik jelöltre, mert szeretik. Nagyjából ugyanennyien azért keresnek rá valamelyik je-

löltre, mert utálják. Mindazonáltal Stuart Gabriellel, a Los Angeles-i Kaliforniai Egyetem pénzügyprofesszorával találtunk egy meglepő nyomot arra vonatkozóan, hogyan terveznek szavazni az emberek. A választással kapcsolatos keresések igen nagy százalékában mindkét jelölt neve szerepel. 2016-ban, a Donald Trump és Hillary Clinton közt zajló választási versengés idején voltak, akik arra kerestek rá, hogy „Trump Clinton közvélemény-kutatás”. Másokat a „Clinton-Trump-vita” fontosabb eseményei érdekelték. Összességében a „Trump” szót tartalmazó keresések 12 százaléka tartalmazta a „Clinton” szót is. A „Clinton” szót tartalmazó kereséseknek több mint negyedében szerepelt a „Trump” szó.

Rájöttünk, hogy ezek a látszólag semleges keresések valójában adhatnak némi támpontot arra vonatkozóan, melyik jelöltet támogatja valaki.

Hogy hogyan? Hát a jelöltek nevének sorrendjével. A kutatásunk szerint az ember nagy valószínűséggel az általa támogatott jelölt nevét írja be először egy olyan keresés során, amely mindkét jelölt nevét tartalmazza.

Az előző három elnökválasztáson az a jelölt kapott több szavazatot, akinek neve több Google-keresésben szerepelt elsőként. És ami még érdekesebb: a jelöltek nevének sorrendje azt is előre jelezte, hogy ki szerez többséget egy-egy államban.

A jelöltek nevének sorrendje a jelek szerint más olyan információt is tartalmaz, amely a közvélemény-kutatók előtt rejtve maradhat. A 2012-es választáson, ahol Obama vetélytársa a republikánus Mitt Romney volt, Nate Silver, a zseniális statisztikus és újságíró mind az ötven államban eltalálta a végeredményt. Mi azonban azt is kiderítettük Stuart Gabriellel, hogy azokban az államokban, ahol többször írták be elsőként Romney nevét, mint Obamáét, Rom-

ney jobban teljesített, mint ahogy azt Nate Silver megjósolta. Azokban az államokban pedig, ahol Obama neve került többször az első helyre, Obama ért el jobb eredményt, mint amit Silver várt.

Ez a mutató azért tartalmazhat olyan információt, amelyre a felmérések során nem derül fény, mert a választók vagy önmaguknak is hazudnak, vagy nem szívesen fedik fel a közvélemény-kutatók előtt a valódi preferenciáikat. Talán azok, akik a 2012-es felmérésekben bizonytalan-nak mondták magukat, de következetesen olyanokat írtak be a keresőbe, hogy „Romney–Obama közvélemény-kutatás”, „Romney–Obama vita” és „Romney–Obama-választás”, mindvégig úgy tervezték, hogy Romney-ra adják majd a voksukat.

Szóval a Google megjósolta Trump győzelmét? Nos, még igen sok a tennivalónk – és rengeteg kutatónak kell még csatlakoznia a munkához – ahhoz, hogy kiderüljön, miként használhatjuk fel a Google-adatokat a választási eredmények előrejelzésére. Ez egy teljesen új tudományág, és csak a legutóbbi néhány elnökválasztásra vonatkozóan rendelkezünk ilyen adatokkal. Semmiképpen nem állítom, hogy már ott tartunk – vagy valaha is ott fogunk tartani –, hogy végleg kidobhatjuk az ablakon a közvélemény-kutatást mint a választási eredmények előrejelzésében hasznos eszközt.

Ugyanakkor az internet számos pontján határozottan voltak előjelek, amelyek arra utaltak, hogy Trump jobban szerepelhet, mint azt a közvélemény-kutatások jósolják.

Az elnökválasztás voltak arra utaló jelek, hogy a választók összetétele kedvezhet Trumpnak. A fekete amerikaiak azt mondták a közvélemény-kutatóknak, hogy hatalmas számban fognak Trump ellen szavazni. Ám a túlnyomórészt feketék lakta területeken igen alacsony volt az olyan

Google-keresések száma, amelyek a szavazással kapcsolatos információkra irányultak. A választás napján Hillary Clinton számára súlyos csapást jelentett a feketék alacsony részvétele.

Még arra utaló jelek is voltak, hogy a bizonytalan szavazók Trump felé mozdulnak. Stuart Gabriellel felfedeztük, hogy a kulcsfontosságú középnyugati államokban, ahol Clinton győzelemre számított, többen kerestek arra, hogy „Trump–Clinton”, mint arra, hogy „Clinton–Trump”. És Trump éppen annak köszönhetette megválasztását, hogy ezekben az államokban jócskán túlteljesítette a közvélemény-kutatói eredményeit.

A sikere legfőbb előjele azonban – már az előválasztás idején – szerintem az a gondosan titkolt rasszizmus volt, amelyre az Obama-tanulmányom fényt derített. A Google-keresések olyan, az amerikaiak jelentős hányadában jelen lévő sötétségről és gyűlöletről tanúskodtak, amely felett a véleményformálók hosszú éveken keresztül elsiklottak. A keresési adatok elárulták, hogy a társadalom, amelyben élünk, erősen különbözik attól, amit a tudósok és az újságírók a közvélemény-kutatói eredmények alapján gondoltak róla. Rusnya, ijesztő, széles körben elharapózott indulatokról árulkodtak az adatok, és ez a felgyülemlett indulat csak egy jelöltre várt, aki majd hangot ad neki.

Az emberek gyakran hazudnak – önmaguknak és másoknak is. Az amerikaiak 2008-ban azt állították a felméréseken, hogy már nem érdekli őket a faji hovatartozás. Nyolc évvel később pedig elnökké választották Donald J. Trumpot, egy olyan embert, aki azt a hamis állítást osztotta meg a Twitteren, hogy a feketék felelősek a fehér amerikaiak ellen elkövetett gyilkosságok többségéért, majd egyik nagygyűlésén védelmébe vette azon támogatóit, akik ellátták a baját a Black Lives Matter Mozgalom egyik

tiltakozó aktivistájának, aztán pedig csak hosszas habozás után vonta meg támogatását a Ku-Klux-Klan egyik egykori vezetőjétől. Ugyanaz a rejtett rasszizmus, amely Barack Obamának hátrányt jelentett, Donald Trumpot segítette.

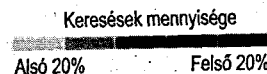
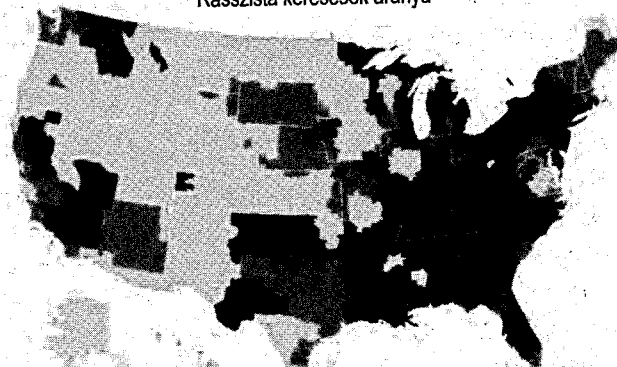
Még az előválasztási időszak kezdetén hangzott el Nate Silver elhíresült kijelentése, mely szerint Trump gyakorlatilag esélytelen. Aztán egyre nyilvánvalóbbá vált, hogy Trump széles körű támogatottságot élvez, és Silver úgy döntött, megnézi az adatokat. Meg akarta érteni, mi folyik itt. Hogy szerepelhet Trump ilyen jól?

Silver észrevette, hogy azok a területek, ahol Trump a legjobb eredményeket érte el, furcsa térképet rajzolnak ki. Trump jól szerepelt az északkeleti országrészben, a Közép-Nyugat iparvidékein, valamint Délen is. Az ország nyugati részén viszont lényegesen rosszabb eredményeket ért el. Silver olyan változókat keresett, amelyek magyarázatot adhatnak erre a térképre. Mi lehetett Trump sikerének titka? A munkanélküliség? A vallásosság? A fegyverbirtoklás? A bevándorlás mértéke? Az Obama-ellenesség?

Silver arra jutott, hogy az az egy tényező, amely a leginkább korrelált Donald Trump támogatottságával a republikánus előválasztáson, az volt, amelyre én bukkantam rá négy évvel korábban. Azokon a területeken támogatták a legtöbben Trumpot, ahol a legtöbbször kerestek rá a „nigger” szóra.

Az utóbbi négy év szinte minden napját Google-adatok elemzésével töltöttem. Egy ideig adattudósként is dolgoztam a Google-nál, ahová azért vettem fel, mert tudomást szereztek a rasszizmussal kapcsolatos kutatásomról. Ezeket az adatokat azóta is folyamatosan vizsgálom a *The New York Times* publicistájaként és adatújságírójaként. És egymást követik a megvilágosodások. Mentális betegségek, emberi szexualitás, gyerekbántalmazás, abortusz,

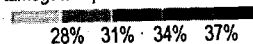
Rasszista keresések aránya



Donald Trump támogatottsága a republikánus előválasztáson



A Trumpot támogató republikánus szavazók becsült aránya



vallás, egészségügy... Nem éppen kis horderejű témák, de a Google-adathalmaz, amely pár évtizede még sehol sem volt, meglepő, új távlatokat nyitott a feltérképezésükben. A közgazdászok és más társadalomtudósok egyfolytában új adatforrásokra vadásznak, úgyhogy engedjék meg, hogy ne kerteljek: meggyőződésem, hogy a Google-keresések képezik az emberi pszichéről valaha összegyűjtött legfontosabb adathalmazt.

Ráadásul ez az adathalmaz nem is az egyetlen eszköz, amit az internetről kaptunk, hogy megérthessük a világot. Hamar rájöttem, hogy vannak más digitális aranybányák is. Letöltöttem az egész Wikipédiát, Facebook-profilokat böngésztem, és a Stormfrontot is megkapirgáltam. Mindezekon felül az egyik legnagyobb pornográf weboldal, a Pornhub is átadta nekem a világ legkülönbözőbb pontjain élő, anonim emberek kereséseire, videómegtekintéseire vonatkozó teljes adatállományát. Fejest ugorhattam abba, amit manapság Big Datának, azaz „nagy adatnak” hívunk. Többtucatnyi emberrel – tudósokkal, szakújságírókkal, vállalkozókkal – készítettem továbbá interjút, akik szintén ennek az új keletű birodalomnak a feltárásán dolgoznak. Számos tanulmányukat fogom tárgyalni ebben a könyvben.

De mindenekelőtt egy vallomás: nem fogom precízen definiálni, hogy mi az a Big Data. Hogy miért? Azért, mert ez eredendően homályos fogalom. Hogyan döntsük el, milyen nagy az, ami nagy? Hol húzzuk meg a határt? Talán ott, hogy 18 462 megfigyelés még Small Data, azaz kis adat, 18 463 pedig már Big Data? A magam részéről inkább minél több mindent lefedő módon tekintek arra, hogy mi minősül ilyennek: az adatok többsége, amelyekkel babrálok, az internetről származik, de más forrásokról is szót ejtek. Napjainkban az elérhető információk mennyiségi és minőségi robbanásának korát éljük, és az új adatok zöme

a Google, illetve a közösségi média révén áramlik be. Az adathalmaz egy része a korábban iratszekrényekben, aktákban rejtőző információk digitalizálásának eredménye. Más része piackutatásokhoz felhasznált, egyre bővülő forrásokból származik. A könyvben tárgyalt tanulmányok némelyike pedig egyáltalán nem dolgozik óriás adathalmazokkal, viszont újszerű, kreatív módon közelíti meg az adatokat – márpedig a mi információktól túltengő körünkben létfontosságú az efféle kreativitás.

Pontosan miért is bír tehát ekkora hatalommal a Big Data? Gondoljunk csak bele, mennyi információ szóródik szét online egy adott napon – ami azt illeti, van is pontos számunk arra vonatkozóan, hogy mennyi. Az emberek a 21. század elején, egy átlagos napon 2,5 millió trillió bájtnyi adatot generálnak.

És ezek a bájtok nyomok.

Egy nő unatkozik csütörtök délután. Rákeres még néhány „nem disznó vicc”-re. Megnézi az e-mailjét. Bejelentkezik a Twitterre. Rákeres a „nigger viccek”-re.

Egy férfi lehangolt. Rákeres arra, hogy „depresszió tünetei”, és arra, hogy „depressziós történetek”. Pasziánszozik egyet.

Egy nő meglátja, hogy egy barátnője bejelentette az eljegyzését a Facebookon. A nő, aki szingli, letiltja a barátnőjét.

Egy férfi kis időre félbehagyja az NFL-lel meg a rappel kapcsolatos böngészést, és feltesz a keresőmotorra egy kérdést: „Normális, ha valaki álmában férfiakkal csókolózik?”

Egy nő rákattint egy BuzzFeed-sztorira, amely bemutatja „a 15 legcukibb cicát”.

Egy férfi is látja ugyanezt a cicás sztorit, de az ő képernyőjén az a felirat, hogy „a 15 legimádnivalóbb cica”. Ő nem kattint rá.

Egy nő rákeres arra, hogy „zseni a fiam?”.

Egy férfi rákeres arra, hogy „hogyan vehetném rá a lányomat, hogy fogjon le?”

Egy nő a hat legjobb barátjával üdül. A barátnői egyfolytában azt mondogatják, mennyire jól érzi magát. Ő pedig titokban rákeres arra, hogy „magány a férjtől távol”.

Egy férfi, az utóbbi nő férje, a hat legjobb barátjával vakációzik. Ő ezt gépeli be titokban a Google-ba: „arra utaló jelek, hogy a feleséged megcsal”.

A fenti adatok némelyike olyan információkat tartalmaz, amit az emberek más körülmények között senkinek sem vallanának be. Ha mindezt összesítjük – természetesen nevek nélkül, hogy soha ne szerezzünk tudomást egyetlen konkrét személy félelméről, vágyairól, viselkedéséről sem –, és beleviszünk a dologba némi adattudományt, máris kezdjük más szemmel látni az embereket: a viselkedésüket, a vágyaikat, a természetüket. Ami azt illeti, megkockáztatva akár a fellengzősség vádját is, arra a meggyőződésre jutottam, hogy a digitalizált korunkban egyre könnyebben hozzáférhető, újfajta adatok radikálisan ki fogják bővíteni az emberiségre vonatkozó tudásunkat. A mikroszkóp megmutatta nekünk, hogy a tó vizének egyetlen cseppjében jóval több van, mint amennyit szabad szemmel látni vélünk benne. A teleszkóp megmutatta, hogy az éjjeli égbolton több a nézni való, mint amennyit szabad szemmel látni vélünk. Az újszerű, digitális adatok pedig most azt mutatják meg, hogy az emberi társadalomban is több van, mint amennyit látni vélünk benne. A Big Data lehet korunk mikroszkópja vagy teleszkópja, amely fontos, sőt forradalmi jelentőségű felismeréseket tesz lehetővé.

Egy ilyen állításban egy további kockázat is rejlik: az, hogy nem pusztán fellengzősen hangzik, hanem még trendi is. Számos nagyotmondó kinyilatkoztatás hangzott már

el a Big Data hatalmára vonatkozóan. De bizonyítékokkal egyik sem nagyon szolgált.

Ez arra inspirálta a „Big Data-szkeptikusokat”, akikből szintén jó néhány akad, hogy elutasítsák a nagyobb volumenű adathalmazok kutatását. Nassim Taleb esszéista és statisztikus például így vélekedik: „Nem mondom, hogy a Big Data nem tartalmaz információt. Van ott információ bőven. Csak az a gond – és ez a középponti kérdés –, hogy egyre nagyobb szénakazalban kell keresnünk a tűt.”

Ennek a könyvnek az az egyik fő célja, hogy felmutassa a hiányzó bizonyítékokat arra vonatkozóan, mit lehet kezdeni a Big Datával – vagyis hogyan találhatjuk meg a tűt az egyre nagyobb szénakazalban. Remélem, sikerül elegendő példával szolgálnom arra, hogy a Big Data újszerű felismerésekkel szolgál az emberi pszichológia és viselkedés terén, hogy végül önök előtt elkezdjenek kirajzolódni valami egészen újszerű, forradalmi dolog körvonalai.

„Várj csak, Seth!” – mondhatják nagyjából most. „Forradalmat ígérsz, és költőien áradozol ezekről a hatalmas, újszerű adathalmazokról. De mindeddig alapvetően két dolgot tudtam meg tőled ezeknek a rendkívüli, elképesztő, lélegzetelállító, úttörő jelentőségű adataidnak a révén: azt, hogy Amerikában sok a rasszista, meg azt, hogy az emberek, különösen a férfiak, tódítanak arról, hogy mennyit szexelnek.”

Beismerem, hogy ezek az újfajta adatok néha csak megerősítik a nyilvánvalót. De aki úgy gondolja, hogy eddig triviális dolgokról esett szó, várjon csak, amíg elér a 4. fejezetig, ahol a Google-keresések alapján egyértelműen, kétségbekonthatatlanul bebizonyítom, hogy a férfiaknál rengeteg aggodalom és bizonytalanság okozója a – dobpergés! – péniszméretük!

Azt állítom, hogy értékes dolog olyasmit bizonyítani, amit eleve gyanítottunk, de nemigen volt rá bizonyítékunk. A gyanú egy dolog. Igazolni már egészen más. De ha a Big Data csak arra lenne jó, hogy beigazolja a sejtéseinket, nem lenne forradalmi jelentőségű. A Big Data azonban, hála az égnek, ennél sokkal többre képes. Az adatok egyre-másra arról árulkodnak nekem, hogy a világ pont fordítva működik, mint gondoltam. Íme, néhány példa, amelyekkel talán sikerül meglepnem önöket.

Talán azt gondolják, hogy a rasszizmus fő oka a gazdasági bizonytalanság és kiszolgáltatottság. Ezért természetes módon arra gyanakszunk, hogy ha az emberek elveszítik a munkahelyüket, a rasszizmus erősödni fog. Pedig valójában sem a rasszista keresések száma, sem a Stormfront taglétszáma nem lesz magasabb, amikor nő a munkanélküliség.

Talán azt gondolják, hogy a szorongás mértéke a túlképzett nagyvárosokban a legmagasabb. A nagyvárosi neurózis közismert sztereotípiája. A szorongással kapcsolatos Google-keresések – úgymint „szorongás tünetei”, illetve „szorongás segítség” – rendszerint mégis gyakoribbak az alacsonyabb iskolázottsági szintű és átlagjövedelmű helyeken, illetve ott, ahol a népesség nagyobb része él vidékies területeken. A szorongás keresési rátája magasabb New York állam vidékies, északi részein, mint New York Cityben.

Talán azt gondolják, hogy a többtucatnyi vagy akár több száz ember életét követelő terrortámadásokat automatikusan nagyarányú, széles körben elharapózó szorongáshullám követi. A terrorizmusnak definíció szerint az a lényege, hogy rettegést váltson ki. Megnéztem a szorongást tükröző Google-kereséseket. Azt vizsgáltam, milyen mértékben nőtt a számuk a 2004 óta végrehajtott nagy európai, illetve amerikai terrorakciók utáni napokban, hetekben, hónapokban. Szóval átlagosan mennyivel is nőtt ezekben

az időszakokban a szorongással kapcsolatos keresések száma? Semennyivel. Egyáltalán semennyivel.

Talán azt gondolják, hogy az emberek olyankor keresnek gyakrabban vicceket, amikor szomorúak. A történelem legnagyobb gondolkodói közül sokan állították, hogy a fájdalomunk enyhítése céljából fordulunk a humor felé. Régóta úgy gondoljuk, hogy a humor a frusztrációval, a fájdalommal, az élet óhatatlan csalódásaival való megbirkózás eszköze. Charlie Chaplin szavaival élve: „A nevetés felüdít, könnyebbséget hoz, és enyhíti a fájdalmat.”

Csakhogy a viccek keresési rátája valójában hétfőnként a legalacsonyabb, pedig az emberek saját bevallásuk szerint olyankor a legboldogtalanabbak. A viccek keresési rátája felhős, esős napokon a legalacsonyabb. Egy-egy komolyabb tragédiát követően pedig a mélybe zuhan, mint például a 2013-as bostoni maraton után, ahol két bombardobbanás során meghalt három ember, és több százan megsebesültek. Az emberek nagyobb valószínűséggel keresnek vicceket a neten, ha szép az élet, és jól mennek a dolgok, mint ha nem.

Egy új adathalmaz néha olyan viselkedésmódokat, vágyakat, aggályokat tár fel, amelyeket eszembe sem jutott volna vizsgálni. Akad jó néhány szexuális hajlam, amely ebbe a kategóriába tartozik. Tudták például, hogy Indiában a leggyakoribb, „a férjem azt akarja, hogy...” kezdetű keresés ez: „a férjem azt akarja, hogy szóptassam”? Ez a komment sokkal gyakoribb Indiában, mint más országokban. Sőt mit több: a férfit szoptató nőt ábrázoló pornóra való keresés négyszer olyan gyakori Indiában és Bangladesben, mint a világ bármely más országában. Ezt egészen biztosan nem gyanítottam, amíg meg nem láttam az adatokat.

Az, hogy a férfiakat rögeszmésen foglalkoztatja a péniszük mérete, nem hangzik túl meglepően, de a nők legna-

gyobb, a testükkel összefüggő és a Google-nak megvallott bizonytalansági tényezője már annál inkább. Ezen új adatok szerint a péniszméret miatti aggodás női megfelelője – rövid hatásszünet – az azon való görcsölés, hogy nem büdös-e a hüvelyük. A nőknél majdnem ugyanannyi a nemi szervük miatti aggályokat tükröző keresés, mint a férfiaknál, csak nekik az a legfőbb gondjuk, hogy milyen a szóban forgó nemi szerv szaga, és miként javíthatnának rajta. Ezt egészen biztosan nem tudtam, amíg meg nem láttam az adatokat.

Az új adatok időnként olyan kulturális különbségekről árulkodnak, amelyek eddig meg sem fordultak a fejekben. Egy példa: az emberek a világ különböző tájain nagyon másként reagálnak arra, hogy a feleségük állapotos. Mexikóban a terhes feleséggel kapcsolatos leggyakoribb keresések között szerepel az, hogy „*frases de amor para mi esposa embarazada*” (szerelmes szavak várandós feleségemnek), valamint az, hogy „*poemas para mi esposa embarazada*” (versek várandós feleségemnek). Az Egyesült Államokban viszont olyanokat találunk a leggyakoribb keresések között, mint hogy „terhes a feleségem, most mi van”, és hogy „terhes a feleségem, mit csináljak”.

A könyvem azonban több mint furcsaságok és egyedi esettanulmányok gyűjteménye, bár lesznek még ilyenek szép számmal. Mivel az alkalmazott módszerek nagyon újak és egyre erőteljesebbek, kifejték néhány gondolatot a működésükről, illetve arról, hogy mitől úttörő jelentőségűek. De a Big Data korlátait is el fogom ismerni.

Az adatforradalomban rejlő lehetőségek iránti lelkesedés időnként félrevisz. A Big Data szerelmeseinek többsége arról ömleng, milyen irdatlan nagyok tudnak lenni ezek az új adathalmazok. Ez az adathalmaz méretével kapcsolatos rögeszme nem újdonság. Dallasban már akkor rendeztek

egy *Nagy és komplex adathalmazok* című konferenciát, amikor a Google, az Amazon és a Facebook még nem volt sehol, sőt még a „Big Data” kifejezés sem létezett. Jerry Friedman, a Stanford statisztikaprofesszora, aki a kollégám volt a Google-nál, felidézte ezt az 1977-es konferenciát. Egy jeles statisztikus emelkedett szólásra. Elmondta, hogy döbbenetes, elképesztő adattömeget halmozott fel: öt gigabájt! Aztán a következő jeles statisztikus emelkedett szólásra. Így kezdte: „Elöttem szóló kollégámnak gigabájtjai vannak. Az semmi. Nekem terabájtjaim vannak.” Más szóval azon volt a hangsúly, hogy mennyi információt tudunk felhalmozni, nem pedig azon, hogy reményeink szerint mit tudunk majd kezdeni vele, illetve milyen kérdéseket tervezünk megválaszolni a segítségükkel. „Akkor mulatságosnak találtam” – mondja Friedman –, „hogy az adatbázisuk méretével akartak lenyűgözni bennünket. Ilyesmi még mindig előfordul.”

Manapság valóban igen sok az olyan adattudós, aki irdatlan méretű adathalmazokat dolgoz fel, amelyekből rendkívül érdektelen dolgokat hoz ki – például hogy a New York Knicks népszerű New Yorkban. Túl sok cég fulladozik az adatokban. Jó sok terabájtjuk van, és annál kevesebb érdemi megfigyelésük. Úgy vélem, az adatbázis méretét sokszor túlértékelik. Van erre egy összetett, de igen fontos magyarázat: minél nagyobb valaminek a hatása, annál kisebb számú megfigyelés kell ahhoz, hogy megértsük. A forró kályhához például elég egyszer hozzányúlunk, rögtön tudni fogjuk, hogy veszélyes. Kávézni viszont több ezerszer kell ahhoz, hogy megállapítsuk, hajlamos-e megfájdulni tőle a fejünk, vagy sem. Vajon melyik lecke a fontosabb? Nyilván a forró kályha leckéje, amely a hatás intenzitása miatt villámgyorsan, minimális adatmennyiség mellett is megmutatkozik.

Az igazság az, hogy a legokosabb Big Data-cégek sokszor tudatosan csökkentik az adatmennységet. A Google-nál a nagy horderejű döntéseket az összes adatból vett parányi minta alapján hozzák meg. Ahhoz, hogy az ember fontos dolgokra jöjjön rá, nincs mindig szükség egy tonna adatra. A megfelelő adatokra van szükség. A Google-keresések hatalmas értékét elsősorban nem az adja, hogy olyan sok van belőlük – hanem az, hogy az emberek annyira őszintén nyilvánulnak meg általuk. Hazudnak a barátaiknak, a szeretőjüknek, az orvosuknak, a közvélemény-kutatóknak és önmaguknak. A Google-lal azonban sokszor kínos információkat is megosztanak, többek között olyan témákban, mint a szexmentes házasságuk, a mentális egészségük, a bizonytalanságaik, vagy éppen a fekete bőrű emberek iránti ellenérzéseik.

És ami a legfontosabb: annak érdekében, hogy jelentős felismeréseket sajtolhassunk ki a Big Datából, a megfelelő kérdéseket kell feltennünk. Ha csak véletlenszerűen pásztázzuk az éjjeli égboltot a teleszkóppal, nem várhatjuk, hogy felfedezze nekünk a Plutót. Ugyanígy nem várhatjuk el egy nagy rakás adattól sem, hogy feltárja előttünk az emberi természet titkait. A kutakodást ígéretes helyeken kell megkezdenünk: például az „a férjem azt akarja, hogy...” kezdetű indiai Google-keresések között.

Ez a könyv azt fogja szemléltetni, hogyan lehet a legjobban használni a Big Datát, és részletesen elmagyarázza, miért is lehet ez hatalmas jelentőségű eszköz. Menet közben az is ki fog derülni, mi mindent felfedeztünk már, én is, mások is, például a következő kérdések kapcsán:

- Hány férfi meleg?
- Működnek-e a hirdetések?
- Miért volt nagy versenyló American Pharoah?
- Elfogult-e a média?

- Létezik-e a freudi elszólás?
- Ki csal adót?
- Számít-e, hogy az ember hova jár egyetemre?
- Kijátszhatjuk a részvényt piacot?
- Hol a legjobb gyereket nevelni?
- Mitől lesz virális egy sztori?
- Miről beszéljünk az első randin, ha azt szeretnénk, hogy második is legyen?

...és még rengeteg, rengeteg további kérdés kapcsán.

De mielőtt mindebbe belevágnánk, meg kell vitatnunk a legalapvetőbb kérdést: azt, hogy miért is van egyáltalán szükségünk adatokra. Ennek érdekében pedig mindennek előtt bemutatom a nagymamámat.

I. RÉSZ

NAGY ADAT, KIS ADAT

Ösztönös „félreérzéseink”

Ha az ember harminchárom éves, és már több, egymást követő hálaadáson partner nélkül jelent meg; valószínűsíthető, hogy szóba kerül a párválasztás témája. Amelyről nagyjából mindenkinek lesz véleménye.

– Sethnek egy bolond lányra van szüksége, olyanra, mint ő – mondja a nővérem.

– Te vagy a bolond! Épp hogy normális lányra van szüksége, aki ellensúlyozza az őrültségét! – mondja a bátyám.

– Seth nem bolond – mondja anyám.

– Te vagy a bolond! Még szép, hogy Seth őrült! – mondja apám.

És ekkor váratlanul megszólal az én szégyellős, halk szavú nagymamám, aki a vacsora alatt mindvégig hallgatott. A harsány, agresszív New York-i hangok elcsöndesednek, és minden szem a rövid, sárga hajú, aprócska idős hölgyre szegeződik, akinek még nem tűnt el egészen a kelet-európai akcentusa.

– Seth, neked egy rendes lányra van szükséged. Nem túl csinosra. De nagyon okosra. Tudjon bánni az emberekkel. Legyen társas lény, hogy azért néha te is csinálj valamit. És humora is legyen, mert neked nagyon jó a humorod.

Hogy miért övezi a családukban ilyen tiszteletteljes figyelem ennek az idős asszonynak a tanácsait? Hát azért, mert az én nyolcvannyolc éves nagymamám többet látott

az életből, mint az asztalnál ülők közül bárki. Több házasságot volt alkalma megfigyelni, mint bármelyikünknek: olyanokat, amelyek működtek, és olyanokat is, amelyek nem. És az évtizedek során katalogizálta a sikeres párkapcsolatot eredményező tulajdonságokat. Így ebben a kérdésben ő fér hozzá közülünk a legtöbb adatponthoz.* A nagymamám a Big Data.

Ebben a könyvben igyekszem varázstalanítani az adattudományt. Akár tetszik, akár nem, tény, hogy az adatok mindannyiunk életében egyre fontosabb szerepet játszanak, és a jelentőségük tovább fog nőni. Az újságok immár terjedelmes rovatokat szentelnek a témának. A cégek komplett teameket tartanak fenn, amelyek kizárólag az adatok elemzésével foglalkoznak. A befektetők dollártízmilliókat tolnak olyan startupokba, amelyek még és még több adatot képesek tárolni. És aki sosem tanul meg regressziót alkalmazni vagy konfidenciaintervallumot számolni, az is rengeteg adattal fog találkozni az általa olvasott weboldalakon, üzleti megbeszéléseken és a vízadagoló automatánál elcsípett pletykákban.

Sokakat szorongással tölt el ez az új fejlemény. Elrettenti őket a rengeteg adat, könnyedén elvesznek és összezavarodnak a számok világában. Úgy vélik, a világ kvantitatív megértése a bal agyféltekés csodagyerekek kiváltsága. Abban a pillanatban, hogy számokba botlanak, legszívesebben máris lapoznának, lefújnak a megbeszélést vagy témát váltanának.

Én azonban tíz évet húztam le az adatelemző üzletágban, és nagy szerencsém volt, mert a szakterület legjobbjaival dolgozhattam együtt. Az egyik legfontosabb lecke,

* Adatpont: a statisztikában bármilyen tény, méréssel megragadható és számokban kifejezhető információ. (A szerk.)

amit megtanultam, a következő: a jó adattudomány nem olyan komplikált, amilyennek az emberek hiszik. Sőt, ami azt illeti, az adattudomány meglepően intuitív tevékenység.

Hogy mitől intuitív tevékenység az adattudomány? Az egész lényege az, hogy szabályszerűségek, „mintázatok” felismeréséről szól, illetve annak előrejelzéséről, hogy miként hat egyik változó a másikra. Az emberek egyfolytában ezt csinálják.

Gondoljunk csak a nagymamám párkapcsolati tanácsaira! Ő is egy hatalmas adatbázist hívott segítségül, amelyet az agya egész hosszú élete során, közel egy évszázad alatt letöltött – családtagoktól, barátoktól, ismerősöktől hallott történetek adatbázisát. Az elemzését szűkebb mintára korlátozta: azokra a kapcsolatokra, amelyekben a férfi hasonló tulajdonságokkal rendelkezett, mint én, vagyis érzékeny, elszigetelődésre hajlamos, humorra fogékony ember volt. Elkülönítette a nő legfontosabb tulajdonságait: azt, hogy mennyire volt kedves, mennyire volt okos, mennyire volt csinos. Ezeket az alapvető kritériumokat párba állította a párkapcsolat legfontosabb kritériumával: hogy jól sikerült-e. Végezetül pedig eredményt hirdetett. Más szóval: mintázatokat ismert fel, és megjósolta, hogyan fog hatni egyik változó a másikra. A nagy adattudós.

Őn is adattudós. Gyerekkorában bizonyára észrevette, hogy ha sír, az édesanyja odafigyel önre. Ez adattudomány. Felnőttként arra lett figyelmes, hogy ha túl sokat panaszkodik, kevésbé akarnak együtt lenni önnel az emberek. Ez is adattudomány. Észrevette továbbá, hogy ha mások kevésbé vágnak a társaságára, akkor ön kevésbé boldog. Amikor kevésbé boldog, kevésbé barátságos. Amikor pedig kevésbé barátságos, akkor még kevésbé vágnak a társaságára. Adattudomány, adattudomány, adattudomány.

Mivel az adattudomány ilyen természetes dolog, tapasztalataim szerint a legjobb Big Data-tanulmányokat nagyjából bárki megértheti, aki általában véve elég okos. Ha ön nem ért meg egy tanulmányt, akkor alighanem a tanulmánnyal van probléma, nem pedig önnel.

Bizonyítékra vágnak, hogy a jó adattudomány rendszert intuitív? Nemrégiben belefutottam egy tanulmányba, amely az utóbbi évek talán legfontosabb ilyen tárgyú munkája, ráadásul az egyik legintuitívabb tanulmány, amelyet valaha láttam. Szeretném, ha nemcsak a jelentőségén tűnődnének el – hanem azon is, hogy mennyire természetes és „nagyis”.

A szóban forgó tanulmányt a Columbia Egyetem és a Microsoft munkatársai jegyzik. A kutatócsoport azt vizsgálta, hogy milyen tünetek jelzik előre a hasnyálmirigyrákot. Ennek a betegségnek alacsony – mindössze 3 százalékos körüli – az öt éves túlélési aránya, de ha korán felismerik, az megduplázhatta a páciens esélyeit.

Milyen módszert alkalmaztak a kutatók? A Microsoft keresőmotorja, a Bing több tízezer anonim felhasználójának adatait használták fel. A közelmúltban hasnyálmirigyrákkal diagnosztizált betegként kódolták* az olyan felhasználókat, akik olyan félreérthetetlen kereséseket futtattak, mint például hogy „most diagnosztizáltak hasnyálmirigyrákkal” vagy „megtudtam, hogy hasnyálmirigyrákom van, mire számíthatok”.

* Kódolás: egy társadalomkutatásban vizsgálatban a releváns információk, nyers adatok standardizálása, mérhetővé tétele. Jelen esetben például a számtalan keresés közül a releváns esetek elkülönítésére alkalmas kódot kapnak a betegségekre utaló keresőkifejezéseket futtató felhasználók, akik így a hasnyálmirigyrákkal diagnosztizált betegek kategóriáját fogják alkotni. (A szerk.)

Ezután az egészségügyi tünetekre irányuló kereséseket nézték meg. Összevetették azt a kevés felhasználót, aki később hasnyálmirigyrák-diagnózisról számolt be, azokkal, akik nem. Vagyis: milyen tünetek jelzik előre, hogy egy felhasználó néhány héttel vagy néhány hónappal később diagnózisról számol be?

Az eredmény döbbenetes volt. Kiderült, hogy a hátfájásra, majd borsárgaságra való keresés a hasnyálmirigyrák előjele; ha viszont valaki csak a hátfájásra keres rá, akkor valószínűtlen, hogy hasnyálmirigyrákja van. Ennek a betegségnek a bizonyítéka az emésztési zavarokra, majd hasfájásra való keresés is, míg ha valaki csak az emésztési zavarokra keres rá, a hasfájásra pedig nem, akkor valószínűtlen, hogy hasnyálmirigyrákja van. A kutatók ezzel a módszerrel az összes eset 5–15 százalékát tudták azonosítani, gyakorlatilag fals pozitív eredmény nélkül. Na most, ez talán nem tűnik valami hatalmas aránynak, de akinek hasnyálmirigyrákja van, az akkor is úgy érzi, hogy kincset talált, ha csak 10 százalékos esélyt kap arra, hogy megduplázza az esélyeit.

A kutatás eredményeit részletező tanulmányt nehéz teljes egészében megértenie annak, aki nem szakember. Rengeteg szakzsargont tartalmaz, például olyanokat, hogy „Kolmogorov–Szmirnov-próba”, amelynek jelentését, bevallom, elfelejtettem. (Valami módszer annak meghatározására, hogy egy modell passzol-e az adatokhoz.)

Figyeljük meg azonban, hogy alapjában véve mennyire természetes és intuitív ez a figyelemre méltó tanulmány! A kutatók orvosi esetek széles spektrumát vizsgálták, és megpróbálták egy bizonyos betegséghez kötni a tüneteket. Tudják, kik alkalmazzák ezt a módszert annak kiderítésére, hogy valaki beteg-e? Férjek, feleségek, anyák, apák, ápolók és orvosok. Ők is a tapasztalatukra, a tudásukra

alapozva próbálják különböző betegségekhez kötni a lázat, a fejfájást, az orrfolyást, a hasfájást. Vagyis a Columbia Egyetem és a Microsoft kutatói úttörő tanulmányt írtak annak a természetes, kézenfekvő módszernek a segítségével, amelyet mindenki alkalmaz, ha diagnózist akar felállítani.

Na de várjunk csak! Lassítsunk itt egy kicsit! Ha a legjobb adattudomány módszertana sokszor természetes és intuitív, ahogy állítom, akkor felmerül egy alapvető kérdés a Big Data értékére vonatkozóan. Ha az emberek természetükből adódóan adattudósok, és az adattudomány intuitív, akkor miért van szükségünk számítógépekre és statisztikai szoftverekre? Minek ide holmi Kolmogorov-Szmirnov-próba? Nem hallgathatunk egyszerűen a megérzéseinkre, ahogy a nagymamám, ahogy a nővérek és az orvosok?

Pontosan erről szól az a vita, amely Malcolm Gladwell *Ösztönösen: a döntésről másképp* című bestsellerének kiadása után erősödött fel. A könyv az emberek ösztönös megérzéseinek csodáját magasztalja. Gladwell történeteinek szereplői pusztán az ösztöneikre hagyatkozva állapítanak meg olyasmiket, hogy hamisítvány-e egy szobor, elrontja-e a teniszező a következő ütést, vagy hogy mennyit hajlandó fizetni a vevő. Az *Ösztönösen* hősei nem futtatnak regressziót, nem számolnak konfidenciaintervallumot, nem végzik el a Kolmogorov-Szmirnov-próbát, a jóslataik mégis figyelemre méltóan pontosak. Sokan intuitív módon egyetértettek az intuíciót védelmébe vevő Gladwell-lel: bíznak a megérzéseikben. Az *Ösztönösen* rajongói minden bizonnyal lelkesen ünneplik az én bölcs nagymamámat, aki mindenféle számítógép nélkül ad párhuzamos tanácsokat. Az e könyvben bemutatott tanulmányokat, amelyeknek egy része a sajátom, talán kevésbé lelkesen ünne-

pelnék, hiszen számítógép bevonásával készültek. Ha a Big Data – mármint nem a nagyfiéle, hanem a számítógépes – forradalom, akkor bizonyítani kell, hogy erősebb, mint az önmagára utalt intuíciónk, amely, ahogy arra Gladwell rámutatott, sokszor rendkívüli dolgokra képes.

A Columbia Egyetem és a Microsoft tanulmánya kristálytiszta példát szolgáltat arra, hogy a rigorózus adattudomány és a számítógépek révén olyasmiket tudhatunk meg, amelyekre a pusztá megérzés soha nem vezetne rá. Továbbá arra is példa, hogy van, amikor számít az adathalmaz mérete. Olykor túl kevés tapasztalattal rendelkezünk ahhoz, hogy megbízhassunk az önmagára utalt intuícióban. Nem túl valószínű, hogy önök – vagy közeli barátai, esetleg családtagjaik – kellően sok hasnyálmirigyrákos beteget láttak volna eddig ahhoz, hogy kiszűrjék: ha az emésztési zavart hasfájás követi, az másra utal, mint az emésztési zavar önmagában. Sőt, mivel a Bing adathalmaza egyre nő, a kutatók szükségszerűen sok-sok olyan rejtett mintázatot fognak észlelni a tünetek előfordulásának időbeli mintázatairól – a hasnyálmirigyrák és más betegségek esetében is –, amelyek az orvosoknak talán nem tűnnének fel.

Sőt mi több: bár az ösztöneink általában véve elég jól eligazítanak a világ működését illetően, azért elég sokszor nem túl precízek. Adatokra van szükségünk, hogy a kép élesedjen. Vegyük például az időjárásnak a hangulatra gyakorolt hatását. Ön alighanem arra tippelne, hogy –10 fokban sűrűbben tör rá a rosszkedv az emberekre, mint 20 fokban. Ez igaz is. De azt már nem feltétlenül tudná megtippelni, hogy mekkora hatása lehet ennek a hőmérsékletkülönbségnek. Megnéztem, hogyan korrelál egy adott terület depresszióval kapcsolatos Google-kereséseinek száma egy sor különféle tényezővel, köztük a gazdasági helyzettel, az iskolázottság szintjével és a templomok láto-

gatottságával. A téli időjárás hatása minden más tényezőét kenterbe verte. Kiderült, hogy meleg éghajlaton, például Honolulu-ban, 40 százalékkal kevesebb a depresszióval kapcsolatos keresés a téli hónapokban, mind hidegebb tájakon, például Chicagóban. De vajon mekkora jelentőséggel bír ez a hatás? Az antidepresszánsok hatékonyságának optimista olvasata szerint a leghatásosabb ilyen gyógyszerek is csak mintegy 20 százalékkal csökkentik a depresszió előfordulásának gyakoriságát. A Google számadatai alapján tehát elmondhatjuk, hogy ha az ember átköltözik Chicagóból Honolulu-ba, az legalább kétszer olyan hatékony gyógyszer a téli melabúra.*

Ösztönös megérzéseink, ha nem vezérli őket gondos számítógépes elemzés, olykor hatalmas bakot lőnek. A saját tapasztalataink és az előítéleteink elvakíthatnak. A nagymamám például a hosszú évtizedek tapasztalatai alapján jobb párkapcsolati tanácsokat ad, mint bárki más a családban, de azért így is vannak vitatható elképzelései arról, hogy mitől lesz tartós egy kapcsolat. Sokszor hangsúlyozza például annak fontosságát, hogy legyenek közös barátaink. Meg van győződve róla, hogy ez kulcsfontosságú tényezője a házasság sikerének: a kellemes, meleg estétet többnyire a férjével, azaz a nagyapámmal és a szomszédok összetartó csapatával töltötte queensi házuk kis hátsó udvarán, a kerti székekben ejtőzve és jóízű pletykálgatással mulatva az időt.

Vállalnom kell azonban annak ódiumát, hogy feláldozom a saját nagymamámat az igazság oltárán: az adattudomány szerint az elmélete téves. Egy informatikus csapat nemrégiben elemezte ki az emberi kapcsolatokról vala-

* A teljes képhez hozzátartozik, hogy röviddel a tanulmány elkészítése után áttettem a székhelyemet Kaliforniából New Yorkba. Adatok segítségével kideríteni, hogy mit kellene tennünk, sokszor gyerekjáték. Tényleg azt tenni már keményebb dió.

ha felhalmozott legnagyobb adathalmazt: a Facebookot. A vizsgálat tárgyát olyan párok nagy mintája képezte, akik adott időpontban „kapcsolatban” voltak. Egy részük bizonyos idő elteltével is „kapcsolatban” maradt, mások viszont „egyedülállóra” módosították a kapcsolati állapotukat. A kutatók azt szűrték le az adatokból, hogy a közös baráti társaság erős előjele annak, hogy a kapcsolat *nem* lesz tartós. Talán azért, mert nem feltétlenül jó dolog, ha minden estét a partnerünkkel és ugyanazzal a kis társasággal töltünk – lehet, hogy éppen az szilárdítja meg a kapcsolatot, ha külön társaságokban is megfordulunk.

Mint látják, az intuíciónk önmagában, amikor sutba vágjuk a számítógépet, és az ösztönünkre hagyatkozunk, olykor elképesztő dolgokat művel – de hibázni is nagyokat tud. A nagymamám például alighanem egy kognitív csapda áldozatává vált, amelybe időnként mindannyian beleesünk: hajlamosak vagyunk eltúlozni a saját tapasztalataink relevanciáját. Az adattudósok kifejezésével élve: *súlyozzuk* az adatainkat, és ennek során túl nagy súlyt helyezünk egy bizonyos adatpontra: önmagunkra.

A nagymamám csak a nagyapámmal és a barátaikkal folytatott esti smúzolásokra koncentrált, így nem szentelt elegendő figyelmet más párokra. Nem gondolkodott el a sógorán és annak feleségén, akik esténként szintén a szűk, összetartó baráti társaságukkal cseverésztek, de sokat veszekedtek, és végül el is váltak. Nem gondolkodott el a szüleim, vagyis a lánya és a veje házasságán sem. A szüleim a szabad estéiken sokszor külön úton járnak – apám jazzklubba vagy meccsre megy a barátaival, anyám pedig étterembe vagy színházba a barátnőivel: és mégis boldog házasságban élnek.

Ha az ösztönünkre hallgatunk, az is megtéveszthet bennünket, hogy mennyire lenyűgöz minket a drámaiság.

Sokszor túlbecsüljük olyasminek a jelentőségét, amiből emlékezetes sztori kerekedik. A felmérések például arról tanúskodnak, hogy az emberek következetesen előbbre helyezik a gyakori halálokok sorában a tornádót, mint az asztmát. Pedig az asztma valójában hetvenszer több ember életét oltja ki. Ám az asztma általi halál nem számít eseménynek – nincs hírértéke. A tornádó okozta haláleseteknek van.

Ha tehát kizárólag arra hagyatkozunk, amit úton-útfélen hallunk, vagy személyesen megtapasztalunk, sokszor tévesen ítéljük meg, hogyan működik a világ. Míg a jó adattudomány módszertana gyakran intuitív, az eredmény sokszor kontrainuitív. Az adattudós fog egy természetes, intuitív gondolati folyamatot – azt, hogy mintázatokat veszünk észre, illetve értelmezünk –, és telepumpálja azokat szteroiddal, amelyek esetleg arra vezethetnek rá bennünket, hogy a világ egészen másként működik, mint gondoltuk. Így történt akkor is, amikor azt vizsgáltam, melyek a sikeres kosárlabda-pályafutás előjelei.

Gyerekkoromban egyetlenegy dologról álmodoztam: arról, hogy ha majd felnövök, közgazdász és adattudós leszek. Dehogy. Csak viccelek. Igazából profi kosaras akartam lenni, de mindenáron, hogy a példaképemnek, Patrick Ewingnak, a New York Knicks sokszoros All Star-csapat-tag centerének a nyomdokaiba léphessek.

Néha azt gyanítom, hogy valójában minden adattudós egy kölyök, aki arra próbál rájönni, miért nem váltak valóra a gyerekkori álmai. Nem meglepő tehát, hogy az utóbbi időben én is lázas nyomozásba kezdtem: arra voltam kíváncsi, mi kell ahhoz, hogy valakiből NBA-játékos legyen. A nyomozás eredménye meglepő volt. Ami azt illeti, újfent azt demonstrálta, hogyan változtathatja meg a jó adattu-

domány a világról alkotott képünket, valamint hogy a számok mennyire kontrainuitívak tudnak lenni.

A következő konkrét kérdésre kerestem a választ: a szegény vagy a középosztálybeli családba született gyerekeknek van-e több esélyük arra, hogy bekerüljenek az NBA-ba?

A többség az előbbire tippelne. A közkeletű vélekedés szerint, ha valaki nehéz körülmények között nő fel, például önkormányzati bérlakásban, egyedülálló, tizenéves anyuka gyermekeként, az elősegíti annak a belső hajtóerőnek a kialakulását, amely elengedhetetlen ahhoz, hogy az ember eljuthasson a csúcsig ebben a kőkemény versenysportban.

Ezt az álláspontot fejtette ki William Ellerbe, a philadelphiai középiskolai kosáredző is a *Sports Illustrated*-nek adott interjújában. „A kertvárosi kölykök többnyire csak szórakozásból játszanak. A belvárosiak számára viszont a kosárlabda élet-halál kérdése” – mondta Ellerbe. Engem sajnos házas szülők neveltek, egy New Jersey-i kertvárosban. Nemzedékem legjobb kosarasát, LeBron Jamest tizenhat éves, egyedülálló anya szülte, és szegénysorban nevelkedett az Ohio állambeli Akronban.

Az általam végzett internetes felmérésből kiderült, hogy az amerikaiak többsége ugyanazt gondolja, mint Ellerbe mester és jómagam: hogy az NBA-játékosok többsége szegénységben nő fel.

Vajon helytálló ez a közkeletű vélekedés?

Vessünk egy pillantást az adatokra! Az NBA-játékosok szocioökonómiai státusára vonatkozóan nem rendelkezünk részletes, átfogó adatforrással. De mi adatnyomozók vagyunk, így különböző forrásokból (basketball-reference.com, ancestry.com, népszámlálási adatok stb.) származó adatok alapján ki tudjuk deríteni, milyen családi háttér segíti elő leginkább az NBA-s pályafutást. A tanulmány, amint azt látni fogják, változatos adatforrásokra támasz-

odik, amelyek között akad kisebb is, nagyobb is, online is, offline is. Bármilyen izgalmasak is az új, digitális adathalmazok, a jó adattudós nem szégyell ásatag, régimódi forrásokhoz nyúlni, ha azoktól remél segítséget. Egy kérdésre úgy kaphatjuk meg a legkönnyebben a helyes választ, ha az összes elérhető adatot kombináljuk.

Az első releváns adat a játékosok születési helye. Megnéztem, hány fekete és hány fehér fiú született az Egyesült Államok megyéiben az 1980-as években. Aztán rögzítettem, hogy közülük hányan jutottak el az NBA-be. Minden megye adatait összevetettem az ottani háztartások átlagjövedelmével. A megyék népességének faji összetételét is számításba vettem, hiszen – ez önmagában is egy teljes könyv témája lehetne –, a fekete fiúk negyvenszer nagyobb valószínűséggel jutnak el az NBA-ig, mint a fehérek.

Az adatok tanúsága szerint a gazdag megyékben született fiúknak lényegesen több esélyük van arra, hogy az NBA-ben játsszanak. Ha például egy fekete fiú az Egyesült Államok egyik leggazdagabb megyéjében születik, kétszer akkora esélye van arra, hogy NBA-s kosaras legyen, mint a legszegényebb megyékben született, szintén fekete fiúknak. A fehér fiúk esetében pedig a leggazdagabb megyék szülöttei 60 százalékos előnyt élveznek a legszegényebbekhez képest.

Mindez tehát a közkeletű vélekedéssel szemben arról tanúskodik, hogy a szegény férfiak valójában alulreprezentáltak az NBA-ban. A felhasznált adatok azonban nem tökéletesek, hiszen az Egyesült Államok számos gazdag megyéjének, így New York megyének (Manhattannek) is vannak szegényebb részei, mint például Harlem. Így hát mégis lehetséges, hogy a nehéz gyerekkor javítja az NBA-kosarassá válás esélyét. Még több nyomra, még több adatra van szükségünk.

Kinyomoztam hát az NBA-játékosok családi hátterét. Az információkat híryanagokból, illetve a közösségi oldalakról szedtem össze. Ez a módszer meglehetősen időigényesnek bizonyult, ezért az elemzést az 1980-as években született afroamerikai NBA-játékosok közül arra a százra korlátoztam, akik a legtöbb pontot dobták. Egy átlagos amerikai fekete férfival összevetve az NBA szupersztárjai 30 százalékkal kisebb valószínűséggel voltak tinédzser vagy férjezetlen anya gyermekei. A legjobb fekete NBA-kosarasok családi háttere tehát arról is árulkodik, hogy a komfortos családi háttér komoly előnyt jelent, ha az ember sikeres akar lenni.

Mindazonáltal sem a megyénkénti születési adatok, sem a játékosok szűk mintájának családi háttere nem szolgáltat kielégítő információt az összes NBA-kosaras gyerekkoráról. Ezért továbbra sem voltam teljesen meggyőződve arról, hogy a kétszülős, középosztálybeli családokból több NBA-sztár kerül ki, mint az egyszülős, szegény családokból. Minél több adattal tudjuk megpakolni a kérdést, annál jobb.

Aztán eszembe jutott még egy adatpont, amely árulkodó nyomokkal szolgálhat egy ember hátterére vonatkozóan. Két közgazdász, Roland Fryer és Steven Levitt egy tanulmányban felvetette, hogy a feketék keresztnéve jelzésértékű a szocioökonómiai hátterük szempontjából. Fryer és Levitt az 1980-as években, Kaliforniában kiállított születési anyakönyvi kivonatokat tanulmányozta. Megállapították, hogy az afroamerikaiak közül a szegény, iskolázatlan, egyedülálló anyák rendszerint másféle neveket adnak a gyermekeiknek, mint a középosztálybeli, iskolázott, házasságban élő szülők.

A jómódúbb családból származó gyerekek nagyobb valószínűséggel kapnak hétköznapi nevet: több köztük

a Kevin, a Chris és a John. A nehezebb körülmények közé, önkormányzati bérlakásokba született gyerekek nagyobb valószínűséggel kapnak egyedi nevet: több köztük a Knowshon, a Uneek és a Breionshay. A szegénységbe született afroamerikai gyerekek majdnem kétszer akkora eséllyel kapnak olyan nevet, mint az adott évben senki más.

Mi a helyzet tehát a fekete NBA-kosarasok keresztnévvel? Úgy hangzanak, mint a középosztálybeli, vagy inkább mint szegény feketék nevei? Ugyanazon időszak vizsgálata alapján kiderült, hogy a kaliforniai születésű NBA-játékosok feleakkora valószínűséggel kaptak egyedi nevet, mint egy átlagos fekete fiú, ami statisztikailag szignifikáns különbség.

Ismer olyan embert, aki szerint az NBA a gettóból felkapaszkodott kölykök ligája? Mondja meg neki, hogy ha legközelebb meccset hallgat a rádióban, hegyezze a fülét! Különösen arra figyeljen, hogy hányszor cselezi ki Russell Dwigthot, hogy aztán Josh égbe nyúló karjai mellett a labdát kérő Kevinnek passzoljon. Ha az NBA valóban hemzsegne a szegény fekete férfiaktól, a közvetítés egészen másképp hangzana: sokkal több olyanféle nevet hallanánk, mint a LeBron.

Immár háromféle bizonyító erejű adathalmazt szedtünk össze: ismerjük a megyénkénti születési adatokat, a legjobb pontszerzők édesanyjának családi állapotát, valamint a játékosok keresztnévét. Nincs tökéletes adatforrás. Az viszont tény, hogy mindhárom halmaz ugyanazt a történetet támasztja alá. A jobb szocioökonomiai státusz nagyobb esélyt jelent arra, hogy az emberből NBA-játékos legyen. Más szóval a közkeletű vélekedés téves.

A nyolcvanas években született afroamerikaiak mintegy 60 százalékanak a szülei nem voltak házasok. Ugyanakkor számításaim szerint azok jelentős többsége, akik be-

kerültek közülük az NBA-ba, házas szülők gyermeke volt. Az NBA játékosállományát tehát elsősorban nem a LeBron Jameshez hasonló hátterű férfiak alkotják. Inkább olyanok, mint Chris Bosh, akit két szülő nevelt fel Texasban, és aki kiélhette az elektronikus kütyük iránti érdeklődését; vagy Chris Paul, aki középosztálybeli szülők második fiaként látta meg a napvilágot az észak-karolinai Lewisville-ben, és akinek a családja 2011-ben szerepelt a *Családi vi-szály* című vetélkedőben.

Az adattudós célja az, hogy megértse a világot. Ha kont-raintuitív eredményt kapunk, még több adattudomány segítségével magyarázatot találhatunk arra, hogy miért nem olyan a világ, amilyennek látszik. Például miért élveznek előnyt a kosárlabdában a középosztálybeli férfiak a szegény férfiakhoz képest? Erre legalább kétféle magyarázat van.

Az első az, hogy a szegény férfiak jellemzően alacsonyabbra nőnek. A tudósok már régen megállapították, hogy a gyermekkori egészségügyi ellátás, illetve táplálkozás komoly szerepet játszik a felnőttkori egészségi állapotunkban. Ebből adódik az, hogy a fejlett világban egy átlagos férfi ma bő tíz centivel magasabb, mint másfél évszázada. Az adatok tanúsága szerint a szegény sorsú amerikaiak alacsonyabb termetűek, mivel életük korai szakaszában rosszabb egészségügyi ellátást kapnak, és rosszabbul táplálkoznak.

Az adatok azt is elárulják, milyen mértékben befolyásolja a testmagasság, hogy valaki eljut-e az NBA-ig. Az intuíciónk kétségkívül azt súgja, hogy a magas termet előnyt jelent egy reménybeli kosárlabdázónál. Csak vessük össze a pályán szaladgáló játékosok jellemző testmagasságát a lelátón ülő szurkolókéval! (Az átlagos NBA-játékos 6 láb 7 hüvelyk, azaz 200,6 centi, míg az átlagos amerikai férfi 5 láb 9 hüvelyk, azaz 175,3 centi magas.)

Mennyit számít a magasság? Az NBA-játékosok olykor föllentenek egy keveset, ha a természetükről kérdezik őket, az amerikai férfiak testmagasságáról pedig nem létezik nyilvánartás. De ha matematikai módszerekkel durván megbecsüljük, hogy nézhet ki a magasság szerinti eloszlás, és megnézzük az NBA saját számadatait, így is könnyen meggyőződhetünk arról, hogy a magasság tényleg hatalmas jelentőséggel bír – sőt, akár még az is kiderülhet, hogy fontosabb szerepet játszik, mint gondoltuk. Számításaim szerint minden egyes hüvelyk (2,54 cm) nagyjából megduplázza annak esélyét, hogy valaki eljusson az NBA-ba. És ez bármilyen testmagasság esetén igaz: egy 5 láb 11 hüvelyk (180,3 cm) magas férfinak kétszer akkora esélye van az NBA-ra, mint egy 5 láb 10 hüvelykes (177,8 cm) férfinak, és egy 6 láb 11 hüvelykes (210,9 cm) férfinak is kétszer akkora az esélye, mint egy 6 láb 10 hüvelykes (208,3 cm) férfinak. Úgy tűnik, hogy a 6 lábnál (182,9 cm) alacsonyabb férfiak közül körülbelül minden kétmilliomodik jut el az NBA-ig. A 7 lábnál (213,4 cm) magasabbak közül – mind a saját becslésem, mind másoké szerint – nagyjából minden ötödik.

Amint az nyilván feltűnt, az adatok világossá teszik, miért is dőltek dugába a kosárlabda-pályafutásról szőtt álmaim. Nem azért, mert kertvárosban nőttem fel, hanem azért, mert mindössze 5 láb 9 hüvelyk magas vagyok, és fehér (arról nem is beszélve, hogy lassú). Ráadásul lusta is vagyok. Továbbá gyenge az állóképességem, borzalmas a dobáspontosságom, és amikor hozzám kerül a labda, olykor pánikroham tör rám.

Annak, hogy a nehéz családi háttérből érkező fiúknak miért olyan nehéz eljutniuk az NBA-ig, van egy második oka is: olykor híján vannak bizonyos társas készségeknek. A közgazdászok több ezer iskolás gyerek adatainak fel-

használásával kimutatták, hogy a kétszülős, középosztálybeli családok nagyobb eséllyel nevelnek másokban bízni képes, fegyelmezett, koncentrálni jól tudó, összeszedett gyerekeket, mint a szegény családok.

Hogyan siklathat ki a társas készségek hiánya egy ígére-tes kosaraskarriert?

Vessünk egy pillantást Doug Wrenn-nek, a kilencvenes évek egyik legtehetségesebb ifjú kosarasának történetére! Jim Calhoun, aki a Connecticuti Egyetemen az edzője volt, és akinek a keze alól több NBA All Star-csapattag került ki, azt állította, hogy Wrenn magasabbra ugrik, mint bárki, akivel valaha dolgozott. Csakhogy Wrenn a háttérből adódóan komoly kihívásokkal küzdött. Egyedülálló anyja nevelte Seattle Blood Alley nevű negyedében, amely az egyik legdurvább környék a városban. Connecticutban folyton összetűzésbe került a környezetével. Gúnyolta a játékostársait, vitatkozott az edzőivel, és a csapat belső szabályzatával szembemenve mindig bő szabású ruhákat hordott. Ráadásul a törvénnyel is meggyűlt a baja: cipőt lopott egy üzletből, aztán pedig balhézott a rendőrökkel. Calhounnak végül elege lett, és kirúgta a csapatból.

Wrenn kapott egy második esélyt a Washingtoni Egyetemen, de az ottani pályafutását is kisiklatta, hogy képtelen volt kijönni másokkal. Addig huzakodott az edzőjével a játékerceken meg a dobásfajtákon, amíg onnan is kirúgták. Az NBA-ból senki sem fedezte fel a tehetségét, így alacsonyabb ligákban kallódott, majd hazaköltözött az anyjához, és végül börtönbe került testi sértésért. 2009-ben így nyilatkozott a *Seattle Times*-nak: „A karrieremnek befellegzett. Az álmaimnak, a terveimnek befellegzett. Doug Wrenn halott. Ennek a kosarasnak, ennek a csávónak annyi. Volt, nincs.” Wrenn-nek nem csupán ahhoz volt meg a tehetsége, hogy

NBA-játékos lehessen, hanem kifejezetten nagy, sőt legendás kosaras is válhatott volna belőle. A temperamentuma miatt viszont még egy egyetemi csapatban sem tudott megmaradni. Ha élete korai szakaszában stabilabbak a körülményei, ő lehetett volna a következő Michael Jordan.

Jordannek persze szintén lenyűgöző volt a súlypontemelkedése. Ráadásul hatalmas egója is volt, és roppant erős versenyszellem fűtötte. A személyisége tehát nem ütött el olyan nagyon a Wrennétől. Jordan is tudott nehéz eset lenni. Tizenkét éves korában például verekedés miatt eltanácsolták az iskolából. De egyvalami biztosan megvolt neki, ami Wrenn életéből hiányzott: a stabil, középosztálybeli neveltetés. Az apja karbantartási vezető volt a General Electricnél, az anyja pedig bankban dolgozott. És mindketten segítettek a karrierje egyengetésében.

Ami azt illeti, Jordan élettörténete tele van olyan sztorikkal, amelyek arról szólnak, miként óvta őt a családja a szépreményű ifjú sportolókat fenyegető csapdaktól. Arra például, hogy kivágták az iskolából, az édesanyja úgy reagált, hogy magával vitte, amikor dolgozni ment. A fiú nem szállhatott ki a bank előtt parkoló kocsiból: ott kellett olvasgatnia, amíg le nem járt a munkaidő. Amikor pedig a Chicago Bulls felfigyelt rá és lehetőséget adott neki, a szülei és a testvérei felváltva látogatták, megóvva mindattól a kísértéstől, amely kéz a kézben jár a hírnévvel és a pénzzel.

Jordan pályafutása nem is úgy ért véget, mint a Wrenné, a *Seattle Times*-ban megjelenő néhány szóval, amit alig olvasott el valaki. Ehelyett a Hírességek Csarnokában szólalt fel a bevétele alkalmából, sok millió tévé néző füle hallatára. Jordan a beszédében elmondta: „Igyekeztem az élet jó dolgaira koncentrálni: tudják, arra, hogy milyennek

látnak az emberek, mennyire tisztellem őket... és hogyan ítél meg a nyilvánosság. Az embernek meg kell állnia egy pillanatra, hogy elgondolkodjék rajta, mit is tesz. És ezt mind a szüleimtől tanultam.”

Az adatok arról árulkodnak, hogy Jordan teljes joggal mondott köszönetet házasságban élő, középosztálybeli szüleinek. Az adatok arról árulkodnak, hogy a kevésbé jó módú sorsú családokban és közösségekben akadnak olyan NBA-szintű tehetségek, akik nem játszanak az NBA-ban. Ezek az emberek rendelkeztek a megfelelő génekkel, és az ambíció sem hiányzott belőlük, de a lelki alkatuk, a vérmérsékletük nem tette lehetővé, hogy kosaras szupersztárokká válhassanak.

És bármit is sugjon az intuíciónk, nem segít, ha valakinek a nyomorúságos körülményei miatt a kosárlabda „élet-halál kérdése”. Remekül illusztrálják ezt az olyan történetek, mint a Doug Wrenné. És az adatok is ezt igazolják.

2013 júniusában LeBron James interjút adott a tévében, miután másodszor is megnyerte a csapatával az NBA-t. (Azóta a harmadik bajnoki címet is begyűjtötte.) „LeBron James vagyok az Ohio állambeli Akronból – közölte. – Annak is a belvárosából. Úgyhogy igazából nem kellene itt lennem.” A Twitter és a többi közösségi oldal izzott a felháborodott kritikáktól. A bírálók nem értették, hogy mondhatja magát hátrányos helyzetűnek egy ilyen rendkívüli tehetség, akit elképesztően fiatalon kiáltottak ki a kosárlabda nagy reménységének. Pedig valóban rontja az ember esélyeit, ha nehéz körülmények között nő fel, legyen bár megáldva akármennyire jó atletikus adottságokkal. Másként fogalmazva: LeBron James sikere még kivételesebb teljesítmény, mint amilyennek elsőre látszik. Az adatok ezt is igazolják.

II. RÉSZ

A BIG DATA HATALMA

Igaza volt-e Freudnak?

Nemrégiben láttam, hogy valaki a „*penistrian*” szóval írt le egy, az utcán gyalogosan közlekedő embert. Értik, ugye? Nem „*pedestrian*”, azaz gyalogos, hanem „*penistrian*”. Ebbe a kifejezésbe az elírások óriási adathalmazában botlottam bele. Tehát valaki meglát egy gyalogost, és leírja a „*penis*”, azaz „pénisz” szót. Ez biztosan jelent valamit, nem igaz?

Nemrég tudomást szereztem egy személyről, aki azt álmodta, hogy banánt eszik, miközben az oltár elé vezeti a menyasszonyát. Ezt egy alkalmazás óriási adathalmazában láttam, amelyben az emberek az álmaikat jegyzik le. Tehát egy férfi arról fantáziál, hogy miközben éppen feleségül vesz egy nőt, fallikus formájú gyümölcsöt majszol. Ez is biztosan jelent valamit, nem igaz?

Igaza volt-e Sigmund Freudnak? A kérdésre adott legőszintébb válasz mindig is egy vállrándítás volt. Ezt az álláspontot Karl Popper osztrák-brit filozófus képviselte a leghatározottabban. Popper ismert állítása szerint Freud elméletei nem falszifikálhatók, vagyis nem cáfolhatók. Nincs rá megfelelő módszer, hogy ellenőrizzük, igazak-e, vagy hamisak.

Freud azt mondhatná, hogy a „*penistrian*” szót leíró személy egy esetlegesen elfojtott szexuális vágyát jelenítette meg. Az illető erre azt felelhetné, nem jelenített ő meg

semmit, és ugyanígy ejthetett volna bármilyen ártatlan gépelési hibát is: írhatta volna például azt, hogy „*pedaltrian*”. Ez afféle „ő ezt mondta, a másik meg azt mondta”-típusú szituáció lenne. Az esküvője alatt banánt evő férfiről Freud pedig azt mondhatná, hogy titkon egy pénisz járt a fejében, ami azt a titkos vágyát tükrözi, hogy inkább egy férfival házasodna össze, mint egy nővel. Az illető úriember erre azt felelhetné, hogy csak véletlenül álmodott banánról, és simán álmodhatta volna azt is, hogy almát eszik, miközben az oltár elé vonul. „Ő ezt mondta, a másik meg azt mondta...” Freud elméletét nem lehet valódi teszt alá vetni.

Egészen mostanáig.

Az adattudomány Freud elméleteinek jó részét cáfolhatóvá teszi – jó néhány híres teóriájának tesztelésére alkalmas. Kezdjük az álmokban megjelenő fallikus szimbólumokkal! Az alkalmazásra feltöltött álmok adathalmazával segítségével könnyedén megfigyelhetjük, milyen gyakran bukkannak fel a fallikus formájú tárgyak. A vizsgálat fókuszálhat például az ételekre, hiszen azokról sokszor álmodunk, és elég sok a fallosz formájú étel (banán, uborka, hot dog stb.). Ezután felmérhetjük azokat a tényezőket, amelyek okai lehetnek annak, hogy bizonyos fajta ételekkel többször álmodunk, mint másokkal: milyen gyakran eszünk ezeket, mennyire finomak a többség szerint, és igen, azt is, hogy fallikus jellegűek-e.

Tesztelhetjük, hogy két, egyformán népszerű étel, amelyek közül az egyik fallosz formájú, más-más mennyiségben jelenik-e meg az álmainkban. Ha kiderül, hogy a fallosz formájú ételekről nem álmodunk nagyobb valószínűséggel, mint más ennivalókról, akkor a fallikus szimbólumok nem jelentős tényezők az álmainkban. Hála a Big Datának, a freudi elmélet ezen állítása valóban cáfolható lehet.

Adataim a Shadow nevű applikációból származnak, amely arra kéri az embereket, hogy jegyezzék le az álmaikat. Azokat az ételeket kódoltam, amelyek álmok tízezerein bukkantak fel.

Összességében miért álmodunk ételekkel? Az elsődleges tényező az, hogy milyen gyakorisággal fogyasztunk egy bizonyos ételt. Az ember által fogyasztott dolgok közül legtöbbször a víz jelenik meg az álmainkban. Az ételek húszas toplistáján szerepel többek közt a csirkehús, a kenyér, a szendvics és a rizs – csupa kifejezetten „nem freudi” táplálék.

A másik fő tényező, amely befolyásolja az álmainkban való előfordulás gyakoriságát, az, hogy mennyire finom egy bizonyos étel az emberek szerint. A toplista élén két, szintén kifejezetten „nem freudi”, ám híresen finom csemege áll: a csoki és a pizza.

Mi hát a helyzet a fallikus formájú ételekkel? Meglepően gyakran lopóznak-e be az álmainkba? Hát nem.

A banán a második leggyakoribb gyümölcs az álmainkban. Ugyanakkor egyúttal ez a második legtöbbet fogyasztott gyümölcs is. Nincs tehát szükségünk Freudra ahhoz, hogy magyarázatot kapjunk a banános álmok gyakoriságára. Az uborka álmaink hetedik leggyakoribb zöldsége – és a hetedik legtöbbet fogyasztott zöldség. Tehát megint csak arra jutunk, hogy az alakjától függetlenül is van magyarázat arra, miért van jelen az agyunkban, amikor alszunk. Sokkal kevesebbszer álmodunk hot dogról, mint hamburgerről. Ez akkor is igaz, ha számításba vesszük, hogy az emberek több burgert esznek, mint hot dogot.

Összegezve: az összes zöldsége és gyümölcsre elvégzett regresszióanalízis (ezzel a módszerrel választják szét a társadalomtudósok a különböző tényezők hatását) alapján

arra jutottam, hogy egy étel falloszra emlékeztető formája nem eredményezi ezek gyakoribb előfordulását az álmainkban, mint amire az adott ételek népszerűsége alapján számíthatnánk. Freud elmélete cáfolható – és, legalábbis az adataim elemzése alapján, téves.

Most pedig elmélkedjünk egy kicsit a freudi elszólásokról! A pszichológus azt a hipotézist állította fel, hogy tévedéseink – azaz nyelvbtlásaink, elírásaink – tudatalatti vágyainkról árulkodnak, amelyek sokszor szexuális jellegűek. Tesztelhetjük ezt a hipotézist a Big Data segítségével? Íme egy módszer: lássuk, hogy tévedéseink – vagyis elszólásaink – mennyire mutatnak kifejezetten „huncut” irányba. Ha igaz, hogy ilyenkor mélyen eltemetett nemi vágyaink szökdösnek elő belőlünk, akkor elszólásainknak aránytalanul sokszor kell tartalmazniuk olyan szavakat, mint például „pénisz”, „fasz” és „szex”.

Pontosan erre voltam kíváncsi, amikor megvizsgáltam a Microsoft kutatói által összeállított, több mint 40 ezer gépelési hibát tartalmazó adathalmazt. Az adathalmaz olyan hibákat összesít, amelyeket az emberek félregépelnek, de rögtön korrigálják is azokat. A több tízezer elírás között bőven akadtak szexuális jellegűek is. Felbukkant például a már említett „*penistrian*”. Olyan is előfordult, hogy valaki „*security*” (biztonság) helyett azt írta, hogy „*sexurity*”, vagy „*rock*” (kő, szikla) helyett azt, hogy „*cock*” (fasz). Ugyanakkor rengeteg ártatlan elírás is volt. Az emberek „*window*” (ablak) helyett azt írták, hogy „*pindow*”, „*vegetable*” (zöldség) helyett azt, hogy „*fegetable*”, „*afternoon*” (délután) helyett azt, hogy „*aftermoon*”, „*refrigerator*” (hűtőszekrény) helyett pedig azt, hogy „*refriderator*”.

Eltért-e hát az átlagtól a szexuális tartalmú elírások száma?

Hogy ezt teszteljem, a Microsoft adathalmazának felhasználásával először azt modelleztem, milyen gyakran cserélnek fel az emberek tévedésből bizonyos betűket. Kiszámoltam, milyen gyakorisággal írnak *t* helyett *s*-t, *g* helyett *h*-t. Ezután létrehoztam egy webrobotot, vagyis internetes botprogramot, amely automatikusan futtat feladatokat a neten. Nevezzük ezt a programot Error Botnak, azaz hibázó botprogramnak. Az Error Bot ugyanolyan gyakorisággal írt *t* helyett *s*-t, mint a Microsoft tanulmánya szerint az emberek. Ugyanolyan gyakorisággal írt *g* helyett *h*-t – és így tovább. A programot lefuttattam azokra a szavakra, amelyeket az emberek a Microsoft adathalmaza szerint elrontottak. Vagyis az én kis botprogramom olyan szavakat próbált helyesen leírni, mint „*pedestrian*”, „*rock*”, „*windows*” és „*refrigerator*”. Csakhogy közben ő is ugyanolyan gyakorisággal írt *r* helyett *t*-t, mint az emberek, amikor például „*rock*” helyett azt írják, hogy „*tock*” (vagyis a „*tik-tak*”-ból a „*tak*”). Ugyanolyan gyakorisággal írt *r* helyett *c*-t, mint az emberek, amikor „*rock*” helyett azt írják, hogy „*cock*”.

Mi derül ki tehát, ha összevetjük az Error Bot és a természetes módon figyelmetlen emberek elírásait? Mire az Error Bot az emberek hibázási szokásai szerint ejtett néhány millió gépelési hibát, számos freudi jellegű mellüetést is elkövetett. „*Seashell*” (tengeri kagyló) helyett például azt írta, hogy „*sexshell*” (szexkagyló), „*lipstick*” (ajakrúzs) helyett azt, hogy „*lipsdick*” (ajakfasz), „*luckiest*” (legszerencsebb) helyett azt, hogy „*fuckiest*” (legelbaszottabb) – és így tovább. Ráadásul – és most jön a lényeg – az Error Bot, amelynek ugyebár nincs tudatalattija, ugyanakkora eséllyel követett el szexuális tartalmú hibákat, mint a hús-vér emberek. Azzal az általunk, társadalomtudósok által na-

gyon kedvelt megszorítással együtt, hogy további kutatásokra van szükség, ez azt jelenti, hogy az emberek nem követnek el több szexuális jellegű betűhibát, mint amennyi véletlenszerűen is várható.

Ahhoz tehát, hogy megértsük, miért írunk olyasmiket, hogy „*penistrian*”, „*sexurity*” vagy „*cock*”, nincs szükség arra, hogy az elírás valamiféle „tiltott gyümölcs”-höz köthet. Nincs szükségünk olyan elméletre, amely szerint a betűtévesszések az ember titkos vágyait jelenítik meg. Ujjaink efféle „elszólásai” tökéletesen magyarázhatók az elgépelések általában jellemző gyakoriságával. Az emberek sokat hibáznak, és a sok hiba közé időnként olyanok is becsúsznak, mint a „*lipsdick*”, a „*fuckiest*” és a „*penistrian*”. Ha egy majom elég sokáig pötyög a billentyűzeten, előbb-utóbb leírja, hogy „lenni, vagy nem lenni”. Ha egy ember elég sokáig pötyög a billentyűzeten, előbb-utóbb leírja, hogy „*penistrian*”.

Freud elmélete, amely szerint az elszólások tudatalatti vágyainkról árulkodnak, tényleg cáfolható – és az adatok általam elvégzett elemzése szerint téves.

A Big Data szerint a banán egyszerűen csak banán, a „*penistrian*” pedig egyszerűen csak a „*pedestrian*” elírása.

Akkor Freud teljesen mellőlött ezekkel az elméleteivel? Azért nem egészen. Amikor először fértem hozzá a Pornhub adataihoz, olyasmire bukkantam, ami legalábbis bizonyos értelemben freudinak tűnt, és revelatív erővel hatott rám. Ami azt illeti, ez volt az egyik legmeglepőbb dolog, amit az adathalmazokban való kutakodás során megtudtam: azoknak az embereknek a megdöbbentően magas száma, akik vérfertőzést megjelenítő tartalmakra kerestek rá.

A Pornhubon, az egyik legnépszerűbb pornóoldalon a férfiak által indított keresések százas toplistáján tizenhat

a vérfertőző témájú videókra irányult. Figyelmeztetem önöket: szókimondó, a nyugalom megzavarására alkalmas tartalom következik. Olyanokra keresnek rá, hogy „fiútestvér-lánytestvér”, „mostohaanya kúr a fiával”, „anya és fia”, „anya kúr a fiával”, illetve „valódi testvérek”. A férfiak vérfertőző tartalmakra irányuló keresései között többségben vannak az anya és fia főszereplésével forgatott jelenetek. És mi a helyzet a nőkkel? A nők által indított Pornhub-keresések százas toplistáján kilenc vérfertőző témájú videókra irányul, és hasonló ábrándokról tanúskodik – attól eltekintve, hogy a szülő, illetve a gyerek neve többnyire felcserélődik. Vagyis a nők többsége apa és lánya vérfertőző aktusát bemutató jelenetekre keres rá.

A fenti adatokban nem nehéz felfedezni Freud Ödipusz-komplexusának legalább a halovány árnyékát. Freud hipotézise szerint az ellenkező nemű szülővel való szexuális együttlét vágya szinte teljesen általánosnak mondható gyermekkorban, később azonban elfojtjuk. Bárcsak megérthette volna a bécsi pszichológus, hogy analitikus szaktudását Pornhub-adatokon is kipróbálhassa, amelyek arról tanúskodnak, hogy az ellenkező nemű szülő iránti érdeklődés a felnőttekben is jelen van – méghozzá igencsak szókimondó módon –, akik nem is igazán fojtják el ezeket.

A Pornhub-adatok persze nem árulják el nekünk, pontosan kiről is fantáziálnak az emberek, amikor ilyen videókat néznek. Tényleg a saját szüleikkel közösjelnek képzeletben? A Google-keresések további nyomokkal szolgálnak, amelyekből kiderül, hogy rengeteg olyan ember akad, aki ilyen vágyakat dédélget.

Vegyük például a „szexelni akarok a...” kezdetű kereséseket! Az összes lehetséges befejezés közül a leggyakoribb az, hogy „az anyámmal”. Összességében az ilyen keresések több mint háromnegyede vérfertőző jellegű. És ennek

magyarázata még csak nem is a konkrét megfogalmazás. A „vonzódok a...” kezdetű keresések között például még inkább dominál a vérfertőző vágyak beismerése. Megengedem, persze – vállalva a rizikót, hogy csalódást okozok Herr Freudnak –, hogy nem különösebben gyakori keresésekről van szó: az Egyesült Államokban évente néhány ezer ember vallja be, hogy vonzódik az édesanyjához. És valaki azt a hírt is közölhetné Herr Freuddal, hogy a Google-keresések, amint arról a könyvben még szó lesz majd, olykor torzítanak, és a tiltott dolgok felé billentik a mérleg nyelvét.

És még így is. Az emberekben sok olyan, „nem helyénvaló” vonzalom él, amelyről azt feltételeztem volna, hogy gyakrabban fordulnak elő a keresések között. A főnök? A beosztott? A tanítvány? A terapeuta? A páciens? A feleségünk legjobb barátnője? A lányunk legjobb barátnője? A feleségünk húga? A legjobb barátunk felesége? Ezek közül egyik vágy megvallása sem vetekedhet az anya iránti vonzaloméval. És mindez, a Pornhub-adatokkal kombinálva, talán tényleg jelent valamit.

Freud azon általános érvényű állítását pedig, hogy az ember szexualitását nagymértékben formálhatják a gyermekkori élmények, más Google- és Pornhub-adatok is alátámasztják, amelyekből kiderül, hogy az emberekben – legalábbis a férfiakban – töménytelen mennyiségű, a gyermekkorhoz köthető ábránd rögzül. A feleségeknek férjükkel kapcsolatos keresései alapján a felnőtt férfiak egyik legjellemzőbb fétise az, hogy bepelenkázzák és megszojtatják őket, különösen, mint arról már szó esett, Indiában. Igen nagy népszerűsége telt szert a rajzfilmpornó is, vagyis a serdülő fiúk kedvelt rajzfilmhőseit ábrázoló szexjelenetek. Vagy vegyük azt, hogy a férfiak milyen foglalkozású nőket keresnek a leggyakrabban a pornóoldalakon. A 18–24

év közötti férfiak a leggyakrabban olyan nőkre keresnek rá, akik bébiszitterként dolgoznak. Ugyanez igaz a 25–64 éves korosztályra. És a 65 éves idősebbekre. És a top négyben mindegyik korosztály esetében ott van a tanárnő és a pomponlány. Nyilvánvaló tehát, hogy a korai életszakasz az átlagnál nagyobb szerepet játszik a felnőttkori szexuális ábrándok alakulásában.

Egyelőre nem tudtam kideríteni eme példátlan méretű adathalmaz segítségével, hogy pontosan miként formálódnak szexuális preferenciáink. De az előttünk álló évtizedekben társadalomtudós kollégáimmal biztosan sikerül új, cáfolható elméleteket gyártanunk a felnőttkori szexualitásról, és ezeket valós adatokkal tudjuk majd tesztelni.

Már most látok néhány alapotívumot, amelyek kétségkívül részét képezik majd az új, adatokra épülő teóriának. Ez nyilvánvalóan más történet lesz, mint Freudnál, aki a gyermekkor, illetve az elfojtás különálló, jól elhatárolható, egyetemes fázisaiban gondolkodott. Ugyanakkor a Pornhub-adatok között történt első puhatolózásom alapján tökéletesen biztos vagyok benne, hogy a felnőttkori szexualitásról kimondott végső ítéletben is szerepel majd néhány, Freud által is hangsúlyozott kulcsmotívum. A gyermekkor mindenképpen komoly szerephez jut majd. Az anyukák is.

Tíz évvel ezelőtt valószínűleg lehetetlen lett volna így ki-analizálni Freudot. Nyolcvan évvel ezelőtt pedig, amikor még élt, egészen biztosan lehetetlen lett volna. Gondoljuk hát végig alaposan, miért is voltak segítségünkre az újfajta adatforrások! Így azt is könnyebben megérthetjük, miért bír olyan óriási hatalommal a Big Data.

Bizonyára emlékeznek rá, hogy korábban azt mondtam: az irdatlan méretű adathegyek nem generálnak automatikusan nagy felismeréseket. Az adatmennyiség önmagában

túlértékelt dolog. De akkor miért is olyan nagy durranás a Big Data? Miért forradalmasítja az önmagunkról kialakított képünket? Azt állítom, hogy a Big Datának négyféle különleges, egyedi képessége van. A fenti Freud-analízis mindkettőt jól szemlélteti.

Legelőször is, talán már feltűnt önöknek, hogy ebben a Freudról szóló eszmefuttatásban roppant komolyan vesszük a pornográfiát. A könyv további részében is gyakran fogjuk hasznát venni pornográf forrásokból származó adatoknak. Némileg meglepő, hogy a szociológusok csak ritkán teszik ezt: a többségük a hagyományos felmérésekből származó adathalmazokban bíz, amelyekre egész karrierjét építette. Pedig ha egy pillanatra elgondolkodunk, be kell látnunk, hogy a pornó széles körű elterjedtsége – valamint az ebből adódó keresési és megtekintési adatok tömege – mérőföldkő az emberi szexualitás megértéséhez vezető úton, amióta... voltaképpen amióta világ a világ. Olyan adathalmazról van itt szó, hogy a láttára Schopenhauer, Nietzsche, Freud és Foucault szájában is összefutott volna a nyál. Az ő életükben ez az adathalmaz még nem létezett. Még néhány évtizede sem létezett. Ma viszont már létezik. Ahogy létezik még sok-sok további, egyedi adatforrás, amelyek a legkülönbözőbb témakörökre nyitnak ablakot: olyan területekre, amelyekről korábban csak sejtéseink lehettek. *A Big Data első különleges képessége az, hogy új típusú adatokkal szolgál.*

A pornóadatok, illetve a Google-keresések adatai nem csupán újszerűek, hanem őszinték is. A predigitális korban az emberek eltitkolták mások elől a kínos gondolataikat. Mások elől most, a digitális korban is eltitkolják ezeket, de az internet elől, illetve bizonyos konkrét webhelyek, például a Google és a Pornhub elől nem, hiszen ott megőrizhetik anonimitásukat. Az említett webhelyek amolyan

digitális igazságszérumként funkcionálnak – ezért tudjuk leleplezni a vérfertőzés iránti széles körű, intenzív érdeklődést. A Big Data segítségével végre azt is láthatjuk, hogy mit akarnak és mit tesznek valójában az emberek – azaz nem csupán azt, hogy elmondásuk szerint mit akarnak és mit tesznek. *A Big Data második különleges képessége az, hogy őszinte adatokkal szolgál.*

És mivel tömérdek adat áll rendelkezésre, egy-egy populáció parányi részhalmazaira vonatkozóan is fontos információkhoz juthatunk. Példának okáért összevethetjük az uborkával és a paradicsommal álmodók számát. *A Big Data harmadik különleges képessége az, hogy az emberi populáció kis részhalmazaira is ráközelíthetünk.*

A Big Datának van még egy lenyűgöző képessége – olyan képessége, amelyet a Freud kapcsán végzett gyorsselemzéshez nem használtam fel, de egy későbbi kutatás során még hasznát vehetem majd. A Big Data villámgyorsan elvégezhető, kontrollált kísérleteket tesz lehetővé. A segítségével az oksági viszonyokat is tesztelhetjük, nem csupán pusztá korrelációkat. Ilyen teszteket egyelőre nagyrészt a vállalati szférában végeznek, de idővel a társadalomtudósok kezében is hatékony eszköz lesz. *A Big Data negyedik különleges képessége az, hogy oksági viszonyokat feltáró kísérleteket tesz lehetővé.*

Most pedig itt az ideje, hogy egyesével szétszálazzuk ezeket a különleges képességeket, és kiderítsük, pontosan miért is annyira fontos a Big Data.

Adatok újragondolva

Minden hónap egy meghatározott péntekének reggelén, hat órakor Manhattan utcáinak nagy része teljesen kihalt. A járda mentén sorakozó üzletek zárva vannak, a bejáratukat acél biztonsági kapu fedi. A lakások ablakai is sötétek, mindenütt néma csönd honol.

A Manhattan déli részén található globális befektetési bank, a Goldman Sachs irodaházának emeletei azonban már ezen a hajnali órán is fényárban úsznak. Javában duruzsolnak a liftek: a cég több ezer alkalmazottja igyekszik mielőbb az íróasztala elé ülni. Reggel hét órára már szinte mindenki a helyén van.

A hónap bármely más napjának ugyanebben az órájában a környéket nem lenne túlzás álmosnak nevezni. De ezen a bizonyos péntek reggelen izgatott, energikus nyüzsgés pezsdíti fel. Ezen a napon egy bizonyos, hamarosan befutó információ igen nagy hatással lesz a tőzsdepiacra.

A közzététele után néhány perccel ez az információ már olvasható lesz a hírportálokon. A közzététele után néhány másodperccel ezt az információt már nagy hangon fogják tárgyalni, vitatni, boncolgatni Goldmannél és a hasonló pénzügyi cégek százainál. A tényleges események nagy része azonban ma már ezredmásodpercek alatt lezajlik a pénzvilágban. A Goldman és más cégek több tízmillió dollárt fizettek azért, hogy hozzáférésük lehessen olyan

száloptikai kábelekhez, amelyek négy röpké ezredmásodperccel (17-ről 13-ra) csökkentették ennek a bizonyos információnak az utazási idejét Chicago és New Jersey között. A cégek algoritmusai szintén ezredmásodpercek alatt olvassák be az adatokat és az ezek nyomán létrejött tranzakciókat. A szóban forgó, kulcsfontosságú adatok közzététele után tehát a piac egy szempillantásnál is rövidebb idő alatt lendül mozgásba.

Vajon miféle adatokról lehet szó? Mi lehet ennyire értékes a Goldman és sok más pénzintézet számára?

A havi munkanélküliségi ráta.

Ezt a számot – amely olyan nagy hatással van a tőzsdepiacra, hogy a pénzintézetek mindent megtesznek azért, hogy a lehető leggyorsabban tudják fogadni és elemezni az adatokat, illetve megtenni az ezeknek megfelelő lépéseket – telefonos felmérés alapján állapítja meg a Munkaügyi Statisztikai Hivatal (MSH), így mire közzéteszik, az adat már háromhetes – vagyis kétmilliárd ezredmásodperces – lesz.

Tekintve, hogy a cégek dollármilliókat költenek arra, hogy egy-egy ezredmásodpercet lefaragjanak az információáramlás időtartamából, igencsak furcsának tűnhet, hogy a kormány ilyen lassan számolja ki a munkanélküliségi rátát.

Alan Kruegernek, amikor 2011-ben, Obama alatt a Gazdasági Tanácsadók Tanácsának (CEA) elnöke lett, éppen az volt az egyik fő célkitűzése, hogy ezek a kritikus számadatok hamarabb kijöhessenek. Nem járt sikerrel. „Az MSH-nak vagy nincsenek meg a szükséges erőforrásai, vagy lera-gadt a huszadik századi gondolkodásmódnál” – vonta le a tanulságot Krueger.

Mivel eléggé nyilvánvaló, hogy a kormány nem fogja egyhamar felpörgetni a tempót, felmerül a kérdés: van-e gyorsabb módja annak, hogy legalább hozzávetőlegesen megbecsülhessük a munkanélküliséggel kapcsolatos sta-

tisztikai adatokat? Ebben a high-tech korszakban – amelyben az emberek minden egyes internetes kattintását rögzítik valahol – valóban heteket kell várunk arra, hogy kiderüljön, hány embernek nincs munkája?

Az egyik lehetséges megoldást Jeremy Ginsberg munkája ihlette, aki korábban mérnökként dolgozott a Google-nál. Ginsberg figyelte rá, hogy a kormány jókora késéssel hozza nyilvánosságra a közegészségügyi adatokat csak úgy, mint a munkanélküliségi rátát. A Járványvédelmi és -Megelőzési Központnak egy teljes hétre van szüksége ahhoz, hogy közlétegye az influenzára vonatkozó adatokat, pedig az orvosoknak és a kórházaknak jól jönne, ha sokkal hamarabb megkapnák az információt.

Ginsberg gyanította, hogy azok, akiket ágynak dönt az influenza, alighanem rákeresnek a betegséggel kapcsolatos tudnivalókra. Vagyis lényegében beszámolnak a Google-nak a tüneteikről. Úgy vélte, hogy az ilyen keresések száma alapján viszonylagos pontossággal mérhető lehet az influenzafertőzöttség aktuális rátája. És valóban, az olyan keresések, mint az „influenza tünetei” vagy az „izomfájdalom” fontos indikátornak bizonyultak a kór terjedési sebességére vonatkozóan.*

Mindeközben a Google mérnökei megalkottak egy Google Correlate nevű szolgáltatást, amely külsős kutatók számára is lehetővé tette, hogy kedvükre kísérletezzenek hasonló elemzésekkel, az egészségügyön kívül egy sor további területen is. A kutatók bevihetnek a Google Correlate-be bármilyen, bizonyos időszakon belül nyomon követett adatsort, és megnézhetik, milyen Google-keresések korrelálnak a leginkább az adathalmazukkal.

* A Google Flu kezdeti verziójának komoly hibái voltak, de a kutatók a közműltban átkalibrálták a modellt, immár több sikerrel.

Mi például Hal Variannel, a Google vezető közgazdászával annak kimutatására használtuk a Google Correlate-et, hogy mely keresések követik a legjobban a lakásárak alakulását. Mint kiderült, emelkedő lakásárak esetén jellemzően olyasmikre keresnek rá az amerikaiak, mint „80/20-as jelzáloghitel”, „új építésű lakóingatlan” vagy „értéknövekedési ráta”. Ha viszont csökkennek a lakásárak, akkor az olyan keresések válnak gyakorivá, mint „hitelhátralék összege alatti eladás”, „víz alatti jelzáloghitel” (amikor a hitelhátralék összege magasabb, mint a ház értéke) vagy „jelzálog adósságelengedés”.

Felhasználhatók-e tehát a Google-keresések a munkanélküliség lakmuszpapírként is csak úgy, mint a lakásárak vagy az influenza esetében? Megállapíthatjuk-e egyszerűen az emberek Google-keresései alapján, hogy közülük hánynak nincs munkája, és képesek vagyunk-e ezt már jóval azelőtt megtenni, hogy a kormányzat összegezné a saját felmérése eredményeit?

Egy nap bevittem az Egyesült Államok 2004 és 2011 közötti munkanélküliségi adatait a Google Correlate-be. A jelzett időszakban több trillió Google-keresést indítottak. Mit gondolnak, melyik függött össze a legszorosabban a munkanélküliséggel? Talán úgy vélik, hogy a „munkaügyi központ”, vagy valami efféle. Nos, ez is gyakori keresés volt, de messze nem a leggyakoribb. Netán az „új munkahelyek”? Erre is valóban sokszor rákerestek, de egyáltalán nem a legtöbbször.

A leggyakoribb keresőkifejezés az általam vizsgált időintervallumban – és ez egyfolytában változik –, a „Slutload” volt. Bizony, a legtöbbször egy pornóoldalra kerestek rá. Ez első pillantásra furcsának tűnhet, de ha belegondolunk, egy munkanélkülinek feltehetőleg rengeteg agyonütni való ideje akad. Sokan otthon üldögélnek egyedül, és unatkoznak.

Egy másik, magas korrelációt mutató – a számítógépes játékok világába vezető – keresés a „Spider Solitaire”, vagyis a „Pókpaszíánsz” volt. Megint csak nem túl meglepő egy rakás olyan embertől, akinek feltehetőleg rengeteg a fölös ideje, hogy így foglalja el magát.

Persze nem állítom egyetlen elemzés alapján, hogy a Slutloadra vagy a pasziánszra való Google-keresések alapján jósolható meg a legjobban a munkanélküliségi ráta. A munkanélküliek hol ilyen, hol olyan módját választják a kikapcsolódásnak (volt olyan időszak, amikor egy másik pornóoldal, a Rawtube került be a legerősebben korreláló keresések közé), és egyik ilyen konkrét keresés sem fedi le a célcsoport többségét. De összességében úgy találtam, hogy a kikapcsolódáshoz kapcsolódó keresések alapján igenis lekövethető a munkanélküliségi ráta – és ez része lesz az előrejelzésére szolgáló legjobb modellnek.

A fenti példa jól illusztrálja a Big Data első különleges képességét: annak újragondolását, hogy mi is minősül adatnak. A Big Data értéke sokszor nem a méretéből adódik, hanem abból, hogy újfajta információkkal szolgál a kutatónak: olyanokkal, amelyeket idáig senki sem gyűjtött.

A Google előtt is voltak hozzáférhető adatok bizonyos szabadidős tevékenységekről (például az eladott mozijegyek száma), amelyek adtak némi támpontot annak megállapításához, mennyi szabadidejük is van az embereknek. Az viszont, hogy megtudhatjuk, mennyit pasziánszoznak és mennyi pornót néznek, teljesen új dolog – és óriási lehetőségek rejlenek benne. Ebben a konkrét esetben például az ilyen adatok segítségével gyorsabban felmérhetjük, mi a helyzet a gazdaságban – legalábbis addig, amíg a kormány nem talál gyorsabb módszert a felmérés lebonyolítására és az adatok összegzésére.

A kaliforniai Mountain View-ban, a Google épületegyüttesében egészen másként zajlik az élet, mint a Goldman Sachs manhattani központjában. A Google irodái reggel kilenckor még szinte üresek. Ha valaki mégis betéved, az alighanem reggelizni jött, hiszen itt ingyen kapja a banános-áfonyás palacsintát, a tojásfehérje-rántottát, a szűrt uborkalét. Az alkalmazottak egy része eleve nincs a városban: külső helyszínekre járnak találkozókra Boulderben vagy Las Vegasban, netán ingyenes sítúrára mentek a Tahoe-tóhoz. A Mountain View-i campuson dél körül dugig vannak a strandröplabdapályák és a füves focipályák. Életem legjobb burritóját a Google mexikói éttermében ettem.

Hogy lehet a világ egyik legnagyobb és legversenyképesebb techcége, legalábbis látszólag, ennyire laza és nagyvonalú? Hát úgy, hogy a Google olyan módon hasznosította a Big Datát a pénz áramlásának automatizált kiépítésére, ahogy senki más. A cég a könyvem létrejöttében is kulcsszerepet játszott, hiszen a Big Data első számú, utcahosszal vezető forrását a Google-keresések jelentik. De arról sem szabad azonban megfeledkeznünk, hogy maga a Google is újfajta adatok összegyűjtésére építette a sikerét.

Aki elég idős ahhoz, hogy már a huszadik században is internetezett, talán még emlékszik az akkori keresőmotorokra, például a MetaCrawlerre, a Lycosra, az AltaVistára, hogy csak néhányat említsek. És talán arra is emlékszik, hogy ezek a keresőmotorok a legnagyobb jóindulattal is mérsékelten voltak megbízhatók. Ha az embernek szerencséje volt, megtalálta, amit keresett. Sokszor viszont nem. Ha például a kilencvenes évek végén beírtuk valamelyik népszerű keresőmotorba, hogy „Bill Clinton”, az első találatok között biztosan felbukkant egy-egy jelentéktelen weboldal, amely annyit közölt, hogy „Bill Clinton egy szar”,

vagy egy olyan, amelynek tartalmát mindössze egy rossz Clinton-vicc jelentette. Vagyis alighanem nem a a legrelevánsabb információkra akadtunk rá az Egyesült Államok akkori elnökéről.

Aztán 1998-ban színre lépett a Google. És a keresési eredményei tagadhatatlanul jobbak voltak, mint bármely vetélytársáéi. Ha valaki 1998-ban begépelte a Google-ba, hogy „Bill Clinton”, feljött az elnök weboldala, a Fehér Ház e-mail-címe, valamint a Clintonról írt legjobb életrajzok, amelyek a neten elérhetők voltak. A Google olyan volt, mint a varázslat.

Vajon mit csináltak másképp a cég alapítói, Sergey Brin és Larry Page?

A többi böngésző egyszerűen azokra a weboldalakra irányította a felhasználót, amelyeken a keresett kifejezés a leggyakrabban fordult elő. Ha tehát valaki Bill Clinton-ról keresett információt, ezek a keresőmotorok átkutatták a teljes világhálót, és kiadták azokat a webhelyeket, ahol a legtöbbször említették Bill Clinton nevét. Ez a rangsorlási rendszer sok szempontból igencsak tökéletlen volt, és könnyen ki lehetett játszani. Egy viccoldal például, amelynek kezdőlapján, valami eldugott helyen ott virított a „Bill Clinton Bill Clinton Bill Clinton Bill Clinton” felirat, előrébb került, mint a Fehér Ház hivatalos honlapja.*

Brin és Page azonban megtalálta a módját annak, hogyan rögzítsen egy újfajta információt, amely sokkal értékeesebb, mint az egyszerű szószámlálás. Arra jöttek rá,

* Ha valaki 1998-ban a Google előtti kor valamelyik népszerű böngésző-jébe beírta, hogy „autók”, a következő pillanatban már térdig gázolt a pornóoldalakban. Ezek a pornóoldalak sokszor úgy verték át a keresőmotort, hogy fehér betűkkel beírták a fehér háttér elé az „autók” szót. Ilyen módon szert is tettek néhány pluszkattintásra olyan emberektől, akik eredetileg kocsit akartak venni, de a pornó elterelte a figyelmüket.

hogy a weboldalak szerzői sokszor linkelnek be olyan honlapokat, amelyek szerintük segíthetnek a tárgyalt téma jobb megértésében. A *The New York Times* például, ha említést tett Bill Clintonról, megadta olvasóinak a lehetőséget, hogy az elnök nevére kattintva eljussanak a Fehér Ház hivatalos honlapjára.

Minden egyes olyan weboldal, amely ilyen linkeket adott meg, bizonyos értelemben véleményt mondott arról, hol található a legjobb információk Bill Clintonról. Brin és Page minden létező témában összesítette ezeket a véleményeket. Mintegy kiszervezte a rangsorolást: rábízta a *The New York Times*ra, sok millió online levelezőlistára, több száz bloggerre és mindenki másra az interneten. Ha egy csomó ember úgy gondolja, hogy Bill Clintonra vonatkozóan a saját hivatalos honlapja a legfontosabb link, akkor alighanem ezt a weboldalt akarják látni azok, akik rákeresnek Bill Clintonra.

Ezek a linkek olyan adathalmazt jelentettek, amely a többi keresőmotor működtetőinek meg sem fordult a fejükben, és amelyek elképesztően hatékonyan tudták előre jelezni, hogy az adott témában mi lehet a leghasznosabb információ. Az itt a lényeg, hogy a Google nem csupán arra koncentrált, hogy több adatot gyűjtsön össze, mint bárki más, hanem *jobb*fajta adatokat talált. A linkelemzésnek köszönhető, hogy a Google 2000-ben, azaz alig két évvel a beindítása után már az internet legnépszerűbb keresőmotorjává nőtte ki magát. Ma pedig Brin és Page összesített vagyona meghaladja a 60 milliárd dollárt.

Ami érvényes a Google-ra, az érvényes bárkire, aki adatok felhasználásával próbálja megérteni a világot. A Big Data-forradalom nem igazán a több és több adat összegyűjtéséről szól. Hanem a megfelelő adatok összegyűjtéséről.

De nem az internet az egyetlen olyan hely, ahol újszerű adatokat gyűjthetünk, és ahol a megfelelő adatok megszerzése világrengető, forradalmi hatású eredményeket hozhat. A könyvem nagyrészt arról szól, hogy miként segíthetik elő a weben fellelhető adatok embertársaink jobb megértését. A most következő alfejezetnek viszont semmi köze a weben fellelhető adatokhoz. Ami azt illeti, az emberekhez sincs semmi köze. De segítségünkre lehet abban, hogy szemléltetni tudjuk ennek a fejezetnek a fő mondanivalóját: azt, hogy az új, nem konvencionális adatok felfoghatatlanul értékesek. Ahogy azokat az alapelveket is, amelyek hozzásegíthetnek a digitális alapú adatforradalom megértéséhez.

A TEST MINT ADAT

2013 nyarán egy átlagon felüli méretű, fekete sörényű, vörösbarna ló New York állam északabbi részén, egy kis istállóban rostokolt: 151 társával együtt az egyéves csikók nagy Saratoga Springs-i árverésére, az augusztusi Fasig-Tipton Select Yearling Sale-re várt. Abban az évben összesen 10 ezer egyéves csikót árvereztek el az országban.

A gazdagok, ha már egy rakás pénzt kiadnak a versenylovukért, maguknak tartják fenn azt a megtiszteltetést, hogy nevet is adhassanak az állatnak. Így a vörösbarna lónak az árverésen részt vevő állatok nagy többségéhez hasonlóan még nem volt neve. Az istállói száma alapján különböztették meg a többiektől: ő volt a 85-ös.

A 85-ös számú csikó azon az árverésen nemigen tűnt ki bármiben is a társai közül. A pedigreje jó volt, de nem kiemelkedően jó. Az apaménje, Pioneerof [*sic*] the Nile élvonalbeli versenyló volt, de az általa nemzett többi csikó-

ról már nem mondható el, hogy sikert sikerre halmoztak volna. Ráadásul a 85-ös külseje is kételyekre adott okot. A bokáján például volt egy horzsolás, amelyet a vásárlók komolyabb sérülés nyomának vélhettek.

A 85-ös akkori tulajdonosa egy egyiptomi sörmágnás volt, Ahmed Zaját, aki azért jött el Saratoga Springsbe, hogy eladja a lovát, és vegyen helyette néhány másikat.

Mint szinte minden lótulajdonos, Zaját is szakértőket fogadott fel, hogy segítsenek neki kiválasztani a megfelelő lovakat. De ő kicsit másféle szakértőkkel dolgozott, mint a többiek. Az efféle árveréseken feltűnő tipikus lószakértők általában középkorú férfiak, akik javarészt Kentuckyból vagy Florida vidékies területeiről érkeztek. Nem túl iskolázottak, de a családi háttérükből adódóan értenek a lóbizniszhez. Zaját szakértői viszont egy EQB nevű kis cégtől érkeztek. Az EQB vezetője nem valami régi vágású lótartó volt. Az EQB vezetője egy Jeff Seder nevű, Philadelphiából való különc volt egy rakás, a Harvardon összegyűjtött diplomával.

Zaját nem először dolgozott az EQB-val, így pontosan tudta, mire számíthat: Seder csapata néhány napig mustárgatja a lovakat, majd ajánlanak neki vagy ötöt, amelyeket megvásárolhat a 85-ös helyett.

Ezúttal azonban más történt. Seder emberei közölték Zajáttal, hogy nem tudják teljesíteni a kérését. Azaz egyszerűen egyetlen lovat sem tudtak neki ajánlani az aznap árverésre bocsátott 151-ből. Ehelyett váratlanul kérlelni kezdték Zajátot: semmilyen körülmények között, véletlenül se adja el a 85-öst! Az a ló, bizonygatta az EQB csapata, nemcsak a Saratoga Springs-i árverés, hanem az év, sőt talán az évtized legjobb lova. „Inkább a házat adja el, de ezt a lovat ne!” – győzködték Zajátot.

A 85-öst másnap minden csinnadratta nélkül, 300 ezer dollárért megvette egy Incardo Bloodstock nevű férfi. Mint

utóbb kiderült a Bloodstock valójában Ahmed Zaját álneve volt. A sörmágnás hallgatott Seder csapatára, és visszavásárolta a saját lovát, amire addig alig akadt példa. (A szabályok nem tették lehetővé, hogy egyszerűen visszaléptesse a 85-öst az árveréstől, ezért volt szükség az álneves tranzakcióra.) A 2013-as Fasig-Tipton Select Yearling Sale-en 62 ló kelt el a 85-ösnél magasabb áron; közülük kettőért több mint egymillió dollárt fizettek.

Zaját három hónap múlva végre talált egy nevet a 85-ösnek: American Pharoah. És másfél évre rá, egy kellemes, 24 fokos szombat estén, New York City egyik elővárosában American Pharoah elnyerte a Triple Crown,* ami akkor már több mint három évtizede egyetlen lónak sem sikerült.

Vajon mit tudott Jeff Seder a 85-ös számú lóról, amit a jelek szerint senki más? És hogy lett ilyen jó szeme a lovakhoz ennek a harvardi öregdiáknak?

Egy perzselő júniusi délutánon találkoztam először az akkor hatvannégy éves Sederrel. American Pharoah Triple Crown-győzelme óta több mint egy év telt el. A floridai Ocalában találkoztunk, ahol javában zajlott a kétéves lovak egyhetes bemutatója, amelynek csúcspontját az aukció jelentette. A rendezvény tehát hasonlított a 2013-as Saratoga Springs-i árverésre, ahol Ahmed Zaját visszavásárolta a saját lovát.

Sedernek dörgedelmes, Mel Brooks-os hangja és dús haja van, és feltűnően ruganyos a járása. Khakiszínű nadrágot viselt nádrágtartóval, a cége logójával díszített fekete inget, a fülében pedig hallókészüléket.

A következő három nap folyamán elmesélte az élettörténetét – és azt is, hogyan vált ilyen jó lószakértővé. Nem

* Triple Crown: az Egyesült Államokban a három rangos lóversenyt, a Kentucky Derbyt, a Preakness Stakeset, és a Belmont Stakeset is megnyerő, hároméves telivéreknak adható cím. (A szerk.)

mondhatnánk, hogy nyilegyenes út vezetett idáig. Seder, miután magna cum laude és Phi Beta Kappa minősítéssel diplomázott a Harvardon, ugyanott elvégezte a jogot, majd üzleti tudományokból is fokozatot szerzett. Huszonhat évesen a Citigroup elemzőjeként dolgozott New York Cityben, de boldogtalan volt, és úgy érezte, kiégett. Aztán egy délután a cég új, Lexington Avenue-i irodaházának előcsarnokában üldögélt, és megakadt a szeme egy hatalmas falfestményen, amely végtelen mezőt ábrázolt. Eszébe jutott, hogy mennyire szereti a vidéket és a lovakat. Miután hazament, a tükör elé állt, és szemügyre vette magát a háromrészes öltönyében. Akkor és ott eldöntötte, hogy ő nem bankárnak való, és nem arra született, hogy New York Cityben élje le az életét. Másnap reggel fel is mondott a Citigroupnál.

Pennsylvania vidékies részére költözött, ahol egy darabig a textiliparban és a sportgyógyászatban dolgozott különböző helyeken, majd végleg, teljes munkaidőben elkötelezte magát a szenvedélyének: hogy megjósolja a versenylovak sikerességét. A lóversenyzésben a számok meglehetősen durvák. Az ezer kétéves csikó közül, amelyeket az ocalai árverésen (az ország egyik legtekintélyesebb aukcióján) bemutatnak, jó, ha öt fog élete során jelentős pénzdíjú versenyt nyerni. Hogy mi lesz a többi 995-tel? Durván egyharmadukról ki fog derülni, hogy lassú. A másik egyharmaduk lesérül, méghozzá a többségük azért, mert a lábai nem bírják a vágózással járó hatalmas terhelést. (Évente több száz ló pusztul el az amerikai versenypályákon, javarészt lábtörés miatt.) A fennmaradó egyharmadot pedig megtámadja egy kór, amit akár „Bartleby-szindrómának” is nevezhetünk. Herman Melville csodálatos novellájában Bartleby, az írnok abbahagyja a munkát, és azontúl a főnöke minden kérésére így felel: „Inkább nem tenném.” A lo-

vaknál is gyakori eset, hogy versenyzői pályafutásuk korai szakaszában mintha rádöbbenének: nem kell futniuk, ha nincs hozzá kedvük. Sokszor előfordul, hogy egy ló jól kezdi a versenyt, de aztán egyszer csak lelassít, vagy teljesen leáll. Miért is rohángászna körbe-körbe teljes sebességgel az ovális pályán, pláne ha még a patája meg a térdízületei megfájdulnak tőle? Úgy dönt, hogy „Inkább nem tenném”. (Az én szívemet minden Bartleby megdobogtatja, legyen az akár ló, akár ember.)

Ha az esélyek ennyire ellene dolgoznak, hogy választhat hát olyan lovat egy tulajdonos, amely aztán nyereséget fog termelni neki? Az emberek hagyományosan úgy gondolják, hogy a pedigré elemzésével tippelhető meg a legjobban, mennyire lesz sikeres egy ló. A szakértő tehát arról ismerszik meg, hogy kapásból eldarál mindent, amit bárki tudni akarhat a ló apjáról, anyjáról, nagyszüleiéről, testvéreiről. Az ügynökök olyanokat mondanak, hogy például egy nagytestű kanca „a szabályoknak megfelelően jutott a méreteihez”, ami annyit jelent, hogy az anyja felmenői között sok volt a nagydarab ló.

Csak hogy van itt egy kis gond. A pedigré, bár kétségkívül számít, csak részben, méghozzá elég kis részben ad magyarázatot egy versenyló sikereire. Vegyük például azoknak a lovaknak a testvéreit, akik elnyerték „Az Év Lov” címet, az egyik legtekintélyesebb éves kitüntetés! Ezek a lovak a lehető legjobb pedigrével rendelkeznek, hiszen a családfájuk megegyezik a világtörténelem legjobb lovaival. Ezzel együtt több mint háromnegyedük soha életében nem nyer komolyabb versenyt. Az adatok arról árulkodnak, hogy a versenysikerek előrejelzésének hagyományos módszerén jócskán van még mit csiszolni.

Valójában persze nem túl meglepő, hogy a pedigré nem annyira hatékony indikátor az előrejelzésben. Gondoljunk

csak bele, hogyan működik ez az embereknél! Képzeljünk el egy NBA-klubtulajdonost, aki tízéves gyerekekből válogatja össze jövődöbéli csapatát, méghozzá a pedigréjük alapján. Ez a bizonyos tulaj nyilván megbízott volna egy ügynököt, hogy vizsgálja meg Earvin Johnson III-at, „Magic” Johnson fiát. Az ügynök pedig ezt jelentette volna: „A fiú méretei egyelőre igen jók, és a szabályoknak megfelelők, a Johnson-vérvonala jellemzőek. A fiú bizonyára remekül fog látni a pályán, önzetlen lesz, nagydarab és gyors. Jó természetű és társaságkedvelő lesz. A járása magabiztos. Kellemes megjelenésű. Remek befektetés!” Sajnálatos módon a tulajdonosnak tizennégy évvel később lenne egy 188 centis (vagyis profi kosarasnak eleve túl alacsony) divatbloggere. A harmadik Earvin Johnsonnak remekül hasznát vehetnék mezek és formaruhák tervezésében, a pályán azonban aligha.

A mi képzeletbeli NBA-tulajunk, aki úgy válogatja össze a csapatát, mint sok lótulajdonos a versenystájlóját, a divatblogger mellett alighanem lecsapna Michael Jordan fiára, Jeffrey és Marcus Jordanre is – és lenne két közepeszerű egyetemi kosarasa. Sok sikert a Cleveland Cavaliers ellen! Utóbbi csapat fő játékosa LeBron James, akinek az anyukája 165,1 centi magas. Vagy képzeljük el egy országot, ahol a pedigréjük alapján választják a vezetőket, és olyan emberek kezébe kerül a kormányrúd, mint George W. Bush! (Bocs, nem tudtam kihagyni.)

A lóügynökök nem csak a pedigré nézik. Elemzik például a kétéves csikók járását, és egyébként is alaposan szemügyre veszik őket. Ocalában órákon át társalogtam különböző ügynökökkel, és ennyi idő bőven elég volt ahhoz, hogy kiderüljön: vajmi kevésbé értenek egyet abban, hogy mit is keresnek voltaképpen.

Ezekhez a kíméletlen ellentmondásokhoz, bizonytalanságokhoz vegyük még hozzá azt a tényt is, hogy egyes lótu-

lajdonosoknak végtelen mennyiségű pénzük van, és máris adott egy olyan piac, amelynek hatékony működése ellen igen komoly tényezők hatnak. Tíz éve például volt egy kétéves ló, a 153-as, amely gyorsabban futott, mint a többi, az ügynökök többsége csodaszépnek látta, valamint csodaszép volt a pedigréje. Olyan felmenőkkel dicsekedhetett, mint Northern Dancer és Secretariat, akik minden idők legnagyobb versenylovai közé tartoznak. A 153-ast egy ír milliárdos és egy dubaji sejk is meg akarta venni. Szabályos licitháború tört ki közöttük, amely igen hamar presztízsháborúvá alakult ált. A többi lótulajdonos döbbenet figyelte, ahogy az ár egyre magasabbra kúszik. A kétéves csikó végül 16 millió dollárért kelt el, messze a legmagasabb összegért, amit egy lóért valaha kifizettek. A 153-as, amely a The Green Monkey nevet kapta, összesen három versenyen futott, és ezeken mindössze 10 ezer dollárt nyert, majd visszavonultatták.

Sedert soha nem értékelték a lovak kiértékelésének hagyományos módszerei. Egyedül az adatok érdekelték. Az volt a terve, hogy feltérképezi a versenylovak különböző tulajdonságait, és megnézi, melyik korrelál a teljesítményükkel. Fontos megjegyeznünk, hogy a tervét tíz évvel a világháló feltalálása előtt dolgozta ki. A stratégiája leginkább mégis az adattudományon alapult. A történetéből levont tanulságoknak bárki hasznát veheti, aki a Big Datával dolgozik.

Seder ténykedése éveken át csak kudarcot és csalódást hozott a konyhára. Mérté például a lovak orrlyukméretét: a világ első, egyben legnagyobb lóorrlyukméret-adathalmazát hozta létre, kiegészítve a lovak által összesen elnyert díjak mennyiségével. De végül arra jutott, hogy az orrlyukméret nem képes előre jelezni egy ló eredményességét.

Seder EKG-vizsgálatot végzett a lovakon, hogy a szívüket vizsgálja, és elpusztult állatok lábait vagdosta le, hogy mérje a gyors összehúzódású izomrostok tömegét. Egyszer lapátot ragadott egy istálló előtt, mert a lovak ürülékének méretére volt kíváncsi: az volt az elmélete, hogy a verseny előtti túlzott súlyvesztés lelassíthatja a paripát. De egyik paraméter sem korrelált a versenysikerekkel.

Tizenkét évig kellett várni az első nagy áttörésre. Seder akkor úgy döntött, hogy a lovak belső szerveinek méretét fogja vizsgálni. Mivel ehhez nem volt meg a technikai felszereltsége, megalkotta magának a saját, hordozható ultrahangkészüléket. Az eredmények figyelemre méltóak voltak. Kiderült, hogy a szív, azon belül is különösen a bal szívkamra mérete nagymértékben előre jelzi a sikeres pályafutást. Egyértelműen ez volt a legeslegfontosabb változó. Egy további belső szerv is fontos szerepet játszik: a lép. Amelyik lónak kicsi a lépe, az gyakorlatilag nem nyer pénzt.

Sedernek volt még egy-két telitalálat. Például több ezer, vágtagzó lovakról készült videót digitalizált, és kiderítette, hogy bizonyos fajta járások nem korrelálnak a jó versenyeredményekkel. Azt is észrevette, hogy vannak kétéves lovak, amelyek 200 méternyi futás után zihálva lélegeznek. Az ilyen lovakért olykor egymillió dollárt is adnak, de Seder adatai arról árulkodtak, hogy nem lesz belőlük semmi. Ezért oda szokta ültetni egy asszisztensét a célvonal közélébe, hogy szűrje ki a zihálókat.

Az ocalai árverésen felvonultatott mintegy ezer ló közül talán ha tíz fog átmenni Seder összes tesztjén. A pedigrét figyelmen kívül hagyja: csak annyiban érdekli, amennyiben az befolyásolja a ló eladási árát. Mint mondta, „A pedigré annyit jelent, hogy a csikónak talán van egy nagyon pici esélye arra, hogy nagy versenyló váljon belőle. De ha

eleve látom rajta, hogy nagy versenyló lesz, akkor mit érdekel engem, honnan került elő?”

Seder egyik este meghívott az ocalai Hilton Hotelbe, a szobájába. Mesélt a gyerekkoráról, a családjáról, a karrierjéről. Fényképeket is mutatott a feleségéről, a lányáról és a fiáról. Elmondta, hogy amikor a három zsidó diák egyikeként beiratkozott a philadelphiai középiskolájába, még csak 147 centi magas volt. (Az egyetemen aztán 175,3 centisre nőtt.) Mesélt a kedvenc lováról, Pinky Pizwaanskiról. Egy melegsoké tiszteletére vásárolta, és róla is nevezte el. Úgy érezte, hogy bár Pinky, a ló nem volt különösebben sikeres, mindig mindent beleadott.

Seder végül megmutatta a 85-ös számú ló adatait tartalmazó dossziét, azaz pályafutása legnagyobb jóslását. Kiadta a titkát? Lehet, de azt mondta, nem érdekli. A titkok őriztetésénél fontosabb neki, hogy bizonyítsa az igazát, hogy megmutassa a világnak: igenis megérte a húszévi lólábvagdosás, trágyalapátolás, ultrahang-barkácsolás.

Íme, a 85-ös néhány adata:

A 85-ös számú ló (később American Pharoah)
Percentilisei* egyéves korában

	Percentilis
Magasság	56
Súly	61
Pedigree	70
Bal szívkamra	99, 61

* Percentilis: a leíró statisztikában az a szám, amelyből megtudjuk, hogy az adott változóról összegyűjtött összes adat hány százaléka alacsonyabb a vizsgált értéknél. Egyszerűbben megfogalmazva: ha a 85-ös magasságának percentilise 56, az azt jelenti, hogy az összes ló 56%-a alacsonyabb nála. (A szerk.)

Ott állt előttem fehéren-feketén, hogy Seder és csapata miért vált annyira a 85-ös megszállottjává. A bal szívkamrája nagyobb volt, mint a többi vizsgált ló 99,61 százaléka!

És nemcsak a bal szívkamrája volt kivételesen nagy, hanem a többi fontos szerve, így a szíve többi része és a lépe is. Seder úgy találta, hogy általánosságban minél nagyobb a bal kamra, annál jobban fog teljesíteni az adott ló a versenyeken. Az ilyen óriási bal szívkamra azonban betegségre is utalhat, ha a többi belső szerv kicsi. De American Pharoah-nak minden fontos szerve átlagon felüli méretű volt, a bal szívkamrája pedig irdatlanul nagy. A 85-ösről tehát az adatai alapján ordított, hogy olyan ló, amilyenből száz-ezer, vagy talán millió közül egy, ha akad.

Mit tanulhatnak az adattudósok Seder projektjéből?

Először is – és talán ez a legfontosabb – azt, hogy ha új adatok révén próbálunk forradalmasítani egy szakterületet, akkor olyan szakterületet válasszunk, ahol a régi módszerek pocskéak. A Seder által leiskolázott, pedigrémániás lőügynökök módszerein például bőven volt mit csiszolni. Akárcsak a szószámlálás-mániás keresőmotorokon, amelyeket aztán a Google iskolázott le.

A Google törekvésének, hogy a keresési adatok segítségével előre jelezze az influenzát, az volt a gyenge pontja, hogy az influenzát az előző heti adatok alapján, egyszerű szezonális kiigazítással is elég jól tudjuk prognosztizálni. Hogy mennyit tesznek hozzá a keresési adatok ehhez az egyszerű, de hatékony modellhez, az még vita tárgya. Véleményem szerint a Google-keresések mérése inkább az olyan betegségek esetében ígér sokat, amelyeknél a már létező adatok gyengébbek. Így hosszú távon értékesebbnek bizonyulhat valami Google STD-féle dolog, amely a ne-

mi úton terjedő betegségeket prognosztizálhatja, mint az influenzát előre jelző Google Flu.

A másik tanulság az, hogy ha az ember megpróbál előre-jelzéseket készíteni, nem kell túl sokat agyálnia azon, hogy miért működik az általa használt modell. Seder sem tudott kimerítő magyarázatot adni arra, hogy miért olyan fontos az eredményesség szempontjából a bal szívkamra. Azt sem igazán tudta megindokolni, miért is olyan értékes szerv a lép. A lócardiológusok és a lóhematológusok egyszer talán majd megoldják ezt a két nagy rejtélyt. Jelen pillanatban azonban ez nem lényeges. Seder ugyanis jóslásban utazik, nem pedig magyarázásban. Márpedig a jóslási bizniszben egyedül azt kell tudni, hogy működik-e valami – és nem azt, hogy miért.

A Walmart például az összes üzlete eladási adatai alapján dönti el, milyen termékeket tegyen ki a polcra. A 2004-es, az ország délkeleti részét sújtó Frances hurrikán előtt például az áruházlánc – teljes joggal – élt a gyanúperrel, hogy ha a közeljövőben vihar fog megtépázni egy várost, az emberek vásárlási szokásai megváltozhatnak. A korábbi hurrikánok idején eladott mennyiségekből próbálták leszűrni, hogy mit keres ilyenkor a nép. És mi volt a válasz? Az epres Pop-Tarts töltött keksz! Ebből a termékből a hurrikánt megelőző napokban hétszer annyi fogy, mint rendesen.

Az elemzést követően epres Pop-Tartsszal megrakott Walmart-kamionok vágtak neki a 95-ös autópályának: vittek a kekszet a hurrikán sújtotta vidék üzleteibe. És a Pop-Tarts tényleg szépen fogyott.

Hogy miért pont a Pop-Tarts? Alighanem azért, mert sem hűtést, sem főzést nem igényel. Hogy miért pont az epres? Halvány fogalmam sincs. De úgy néz ki, ha hurrikán készül, az emberek az epres Pop-Tartsból merítenek vigaszt. Így aztán a Walmart ma már minden hurrikán előtt feltölti

az üzletek polcait epres Pop-Tartsszal. Hogy miért függ össze a két dolog, az nem számít. Maga az összefüggés viszont igen. Az élelmiszer-tudományi szakemberek egyszer talán majd megfejtik, pontosan mi is a kapcsolat a hurrikán és a töltött, édes keksz között. De addig is, amíg a magyarázatra vár, a Walmartnak jól be kell spájzolnia epres Pop-Tartsból, ha hurrikán közeleg, a Rice Krispiest, vagyis a ropogós rizspelyhet pedig el kell tennie szebb napokra.

Orley Ashenfelter történetének is világos a tanulsága. Ashenfelter, a Princeton közgazdásza nagyjából ugyanaz a borászatban, mint Jeff Seder a lóbizniszben.

Valamivel több mint egy évtizeddel ezelőtt Ashenfelter épp nagyon bosszús volt. Akkoriban rengeteg vörösbort vásárolt Bordeaux környékéről. A bor sokszor fénséges volt, és bőven megérte a borsos árat. Máskor viszont egyenesen kiábrándító volt.

Ashenfelter azon töprengett, vajon miért fizet mindig ugyanannyit az igencsak változó minőségű borért.

Aztán kapott egy tippet egy újságíró barátjától, aki borszakértő is volt. Mint kiderült, valóban van rá mód, hogy megtudjuk, jó lesz-e egy bizonyos bor. Ashenfelter barátja elmondta: a titok nyitja az, hogy milyen az idő a szőlő növekedési időszakában.

Ez felpiszkálta Ashenfelter érdeklődését. Utánajárt, hogy tényleg igaz-e, és van-e rá lehetősége, hogy a jövőben következetesen csak jó borokat vásároljon. Letöltötte a bordeaux-i borvidék harminc évének időjárási adatait. Szintén összegyűjtötte egy csomó bor aukciós árát. Az elárverezett bordói borok ára; mivel az aukcióra évekkal az eredeti eladás után kerül sor, elárulja, mennyire sikerült jól a bor.

Az eredmény elképesztő volt. A bor minőségét igen nagy százalékban magyarázta, hogy milyen volt az idő a szőlő növekedési időszakában.

Ami azt illeti, egy bor minősége felbontható egy egyszerű képletre, amelyet a borászat első törvényének nevezhetünk:

$$\text{Ár} = 12,145 \text{ USD} + 0,00117 \text{ USD} \times \text{téli csapadék cm-ben} + 0,0614 \text{ USD} \times \text{átlaghőmérséklet a szőlő növekedési időszakában} - 0,00386 \text{ USD} \times \text{csapadékmennyiség szüret idején}.$$

De vajon miért így alakul a borminőség a bordeaux-i borvidéken? Mivel magyarázható a borászat első törvénye? Ashenfelter képletére van valamiféle általános magyarázat: napsütésre és korai öntözésre is szükség van ahhoz, hogy a szőlőszemek rendesen beérjenek.

Am a jósképlet pontos számai már nem magyarázhatók semmiféle elmélettel, és az összes részletet valószínűleg a téma szakértői sem fogják soha megérteni.

Miért pont 0,1 centet tesz hozzá az árhoz átlagosan a téli eső minden egyes centiméterére? Miért nem 0,2 vagy 0,05 centet? Erre senki sem tudja a választ. De tény, hogy ha a télen 1000 centiméterrel több eső esik, akkor fel kell készülnünk rá, hogy egy dollárral többet kell fizetnünk egy üveg borért.

Ashenfelter sem tudta pontosan, miért úgy működik a regressziója, ahogy, de aszerint vásárolt borokat, és mint elmondta: „Remekül bevált a dolog.” Az elfogyasztott borok minősége észrevehetően javult.

Ha az a cél, hogy megjövendöljük, melyik bor lesz finom, melyik termék fog szépen fogyni vagy melyik ló lesz a leggyorsabb, nem kell túl sokat agyalnunk azon, hogy miért pont úgy működik a modellünk, ahogy. Az a lényeg, hogy stimmeljenek a számok. Ez Jeff Seder lovas sztorijának második tanulsága.

Az utolsó tanulság az, hogy ha prognosztizálni akarjuk, kiből lehet Triple Crown-győztes, akkor nyitottan, rugalmasan kell meghatároznunk, hogy mi minősül adatnak. Nem mondhatnánk, hogy Jeff Seder színre lépése előtt a régi vágású lóügynökök ne tudták volna, mi az az adat. Buzgón tanulmányozták az időeredményeket és a lovak családfáját. Seder zsenialitása abban rejlett, hogy ott keresett adatokat, ahol mások nem, és a hagyományostól eltérő adatforrásokat is figyelembe vett. Egy adattudósnak nagy hasznára válhat egy friss, eredeti nézőpont.

A SZÓ MINT ADAT

Még 2004-ben történt, hogy két, médiára szakosodott ifjú közgazdász, akik a Harvard PhD-hallgatói voltak, egy friss bírósági döntésről olvasott, amely legalizálta Massachusettsben a melegházasságot.

A két közgazdász, Matt Gentzkow és Jesse Shapiro észrevett valami érdekeset: azt, hogy noha ugyanarról a hírről számolnak be, két különböző újság szóhasználata döbbenetesen eltér egymástól. A konzervatív lap hírében álló *Washington Times* szalagcíme ez volt: „Massachusettsben »összeházasodhatnak« a homoszexuálisok.” A liberálisnak számító *The Washington Post* azt közölte olvasóival, hogy győzelmet arattak az „azonos nemű párok”.

Az nem meglepő, hogy a különböző hírszolgáltatók más-más irányba hajlanak, illetve hogy két különböző újság másra fókuszál ugyanazon sztori tálalásakor. Gentzkow és Shapiro akkor már évek óta azon törte a fejét, miként vehetnék hasznát közgazdasági képzettségüknek a média elfogultságának feltárásában. Miért tűnik úgy, hogy

egyes hírszolgáltatók liberálisabb, mások pedig konzervatívabb álláspontra helyezkednek?

De Gentzkow-nak és Shapirónak nem igazán támadt ötlete a kérdés megoldására: nem tudtak rájönni, miként mérhetnék redszerezetten és objektíven a média szubjektivitását.

Gentzkow és Shapiro akkor, a melegházasságról szóló hír kapcsán nem azt találta igazán érdekesnek, hogy a hírszolgáltatók másként számolnak be ugyanarról; hanem azt, hogy *miben* különböznek a tudósítások. A kulcs a szóhasználat különbözőségében rejlett. A *Washington Times* által használt „homoszexuális” szó 2004-ben már régimódinak és pejoratívnak számított a melegek megjelölésére, a *The Washington Post*-ban olvasható „azonos nemű párok” kifejezés viszont azt hangsúlyozta, hogy a meleg párkapcsolat csupán egy a szerelem sok formája közül.

A két ifjú tudós azon töprengett, vajon a nyelvhasználatban rejlik-e az elfogultság megfejtésének kulcsa. Lehet, hogy a liberálisok és a konzervatívok következetesen használnak más-más szavakat? Adatokká alakíthatók-e az újságok tudósításaiban használt szavak? Mit árulhat el ez az amerikai sajtóról? Kideríthetjük-e, hogy liberális vagy konzervatív-e a sajtó? És kideríthetjük-e, hogy miért? Az ilyen kérdéseken való rágódás 2004-ben már nem számított léha, céltalan időtöltésnek. Az amerikai újságokat megtöltő sok milliárd szó már nem sínylődött a papír, illetve a mikrofilm fogságában. Már voltak bizonyos weboldalak, amelyek szóról szóra rögzítették az Egyesült Államok szinte minden újságjának összes híryanagát. Gentzkow és Shapiro megkapargathatta ezeket a weboldalakokat, és gyorsan letehetelhet, mennyire mérhető az újságok elfogultsága a szóhasználatuk alapján. És ezáltal elmélyíthették a tudásunkat arról, miként is működik a hírmédia.

De mielőtt beszámolnék az eredményeikről, egy pillanatra hagyjuk magára a sajtó nyelvének számszerűsítésén fáradozó Gentzkow-t és Shapirót, és ejtsünk néhány szót arról, hogy miként vették hasznát a legkülönbözőbb tudományágak művelői, az emberi természet kutatói ennek az új adattípusnak: a szavaknak.

A nyelv természetesen mindig is érdekelte a társadalomtudósokat. Csakhogy a nyelv tanulmányozásának előfeltétele általában az volt, hogy az ember átrágja magát mindenféle hosszú szövegen, az irdatlan szövegmenyiség adatokká alakítása pedig gyakorlatilag lehetetlen feladat volt. Ma viszont, a számítógépek és a digitalizáció korában könnyen csoportosíthatjuk a dokumentumok tömkelegében előforduló szavakat. A nyelv ilyenformán a Big Data-elemzés tárgyává vált. A Google által felhasznált linkek szavakból álltak. Ugyanígy az általam tanulmányozott Google-keresések. Ebben a könyvben is sokszor jutnak főszerephez a szavak. A nyelv olyan fontos eleme a Big Data-forradalomnak, hogy megér egy külön részt. Sőt, ma már olyan nagy hasznát veszik a Big Data-kutatások során, hogy már különálló szakterület foglalkozik kifejezetten „a szöveggel mint adattal”.

E szakterület egyik legfontosabb vívmánya a Google Ngrams. Pár évvel ezelőtt két fiatal biológus, Erez Aiden és Jean-Baptiste Michel régi, poros szövegek szavait számoltatta össze a kutatóasszisztenseivel, hogy képet kapjanak bizonyos szavak használatának elterjedéséről. Aztán a fülükbe jutott, hogy a Google új projekten dolgozik: digitalizálja a világ könyveinek jelentős hányadát. A két biológus hamar rádöbbsent, hogy így sokkal könnyebb lenne tanulmányozni a nyelv történetét.

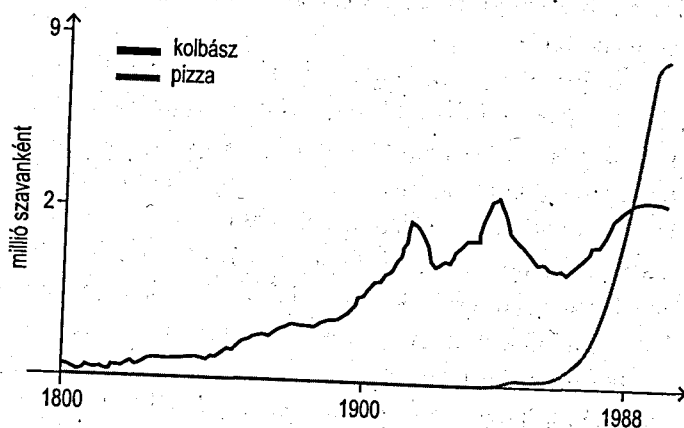
„Beláttuk, hogy a módszereink reménytelenül elavultak” –

mondta a *Discover* magazinnak Erez Aiden. „Világos volt, hogy a digitalizációval nem vehetjük fel a versenyt.” Elhatározták tehát, hogy együttműködnek a keresőcéggel. A Google mérnökeinek segítségével létrehozta egy olyan szolgáltatást, amely digitalizált könyvek millióiban rá tud keresni egy-egy konkrét szóra vagy kifejezésre. A kutatók így kikereshetik, hogy milyen gyakran fordult elő az adott szó vagy kifejezés bármely tetszőleges, 1800 és 2010 közötti évben.

De vajon mit árul el nekünk az, hogy hányszor tartalmaznak egy-egy szót vagy kifejezést egy bizonyos év könyvei? Nos, sokat megtudhatunk például a kolbász népszerűségének lassú növekedéséről, valamint a pizza közkedveltségében beállt villámgyors ugrásról a közelmúltban.

De ez a módszer ennél lényegesebb tanulságokkal is szolgál: a Google Ngrams segítségével többek között azt is megtudhatjuk, hogyan alakult ki a nemzeti identitás. Aiden és Michel *Uncharted (Feltérképezetlen)* című könyve bemutat egy lenyűgöző példát.

De mindenekelőtt egy gyors kérdés: ön szerint az Egyesült Államok egységes vagy megosztott ország? Ha ön is úgy vélekedik, mint a nagy többség, akkor azt mondja erre,

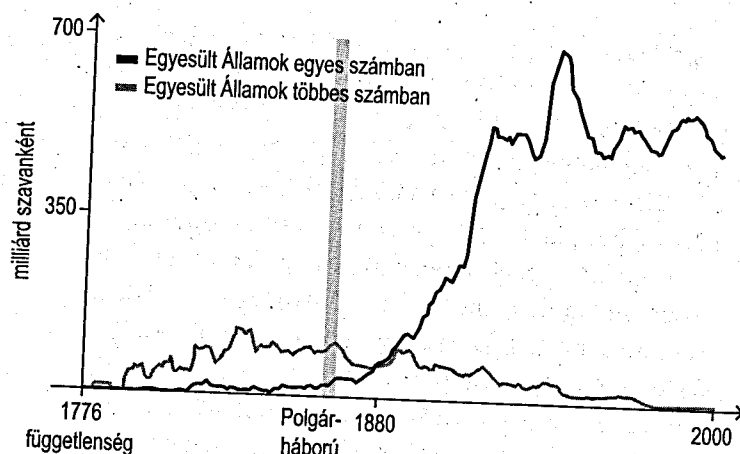


hogy Amerika a politikai polarizáció magas szintje miatt mostanában eléggé megosztott. Sőt, akár azt is mondhatja, hogy megosztottabb, mint valaha. Végül is Amerika ma már „színkódolt” ország: a piros államok republikánusok; a kékek demokraták. De Aiden és Michel *Uncharted* című könyvükben felhívják a figyelmet egy lenyűgöző adatpontra, amely felfedi, hogy az Egyesült Államok sokkal megosztottabb volt egykor, mint ma. A szóban forgó adatpont az a szóhasználat, amellyel az emberek az országról beszélnek.

Idézzük csak fel, milyen szavakat használtam az előző bekezdésben, amikor arról beszéltem, milyen megosztott az ország! Azt írtam: „az Egyesült Államok megosztott”. Egyes számú alanyként hivatkoztam az Egyesült Államokra. Ez magától értetődő: ez a bevett, nyelvtanilag helyes alak. Biztosra veszem, hogy fel sem tűnt senkinek.

Pedig az amerikaiak nem mindig beszéltek így. Az ország történetének korai szakaszában még többes számban emlegették az Egyesült Államokat. John Adams elnök például 1799-es évértékelő beszédében még ezt mondta: „Az Egyesült Államok az Ő Brit Felségével kötött egyezményükben...” Ha 1800-ban írtam volna meg ezt a könyvet, én is úgy fogalmaztam volna, hogy az Egyesült Államok megosztottak. A történészek régóta elalélnek ettől az apró nyelvhasználati differenciától, mivel arra utal, hogy volt egy pont, amelytől kezdve Amerika már nem államok csoportjának, hanem egy nemzetnek tekintette magát.

Mikor is történt ez? Az *Uncharted* című könyvből arról értesülhetünk, hogy a történészek ezt soha nem tudták biztosan, mivel állt módjukban, hogy rendszerezetten vizsgálják a kérdést. De azért sokan gyanították, hogy a polgárháború lehetett a változás oka. Az Amerikai Történeti Társaság volt elnöke, a Pulitzer-díjas James McPherson például kereken kijelentette, hogy „A háború jelentette



az átmenetet: akkor lett az Egyesült Államokból egyes számú főnév”.

De kiderült, hogy McPherson tévedett. A Google Ngrams lehetőséget adott Aidennek és Michelnek, hogy tesztelje ezt a feltevést. Évenkénti bontásban megnézhatték, hogy a története során milyen gyakran említették az Egyesült Államokat egyes, illetve többes számban a könyvek. Az egyes számra való áttérés fokozatosnak bizonyult, és csak jóval a polgárháború után gyorsult fel.

Tizenöt évvel a polgárháború után még mindig gyakrabban említették többes számban az Egyesült Államokat, mint egyes számban, ami arról tanúskodik, hogy nyelvhasználati szempontból még akkor is megosztott volt az ország. A katonai győzelmekhez jóval kevesebb időre van szükség, mint a gondolkodásmód megváltozásához.

Egyelőre ennyit arról, hogy miként egyesül egy ország. De miként egyesül egy férfi és egy nő? A szavak ennek a kérdésnek a megválaszolásában is segítségünkre lehetnek.

Például megjósolhatjuk, hogy egy férfinak és egy nőnek lesz-e második randija is, az alapján ahogy az elsőn beszélgettek egymással.

Ezt a Stanford és a Northwestern tudosaiból álló interdiszciplináris kutatócsoport mutatta ki, név szerint Daniel McFarland, Dan Jurafsky és Craig Rawlings. Több száz heteroszexuális villámrandizót vizsgáltak, és megpróbálták kideríteni, hogyan lehet előre jelezni, hogy két ember úgy fogja-e érezni, összeköti őket valami, illetve hogy szeretnék-e majd másodszor is találkozni.

A kutatók először hagyományos adatokkal próbálkoztak. Azt vizsgálták, mennyire korrelál a villámrandizók magassága, testsúlya, hobbija azzal, hogy ki számol be a randi után romantikus szikráról. A nők általánosságban a magasabb férfiakat részesítik előnyben, illetve azokat, akikkel van közös hobbiuk; a férfiak általánosságban a vékonyabb nőket részesítik előnyben, illetve azokat, akikkel van közös hobbiuk. Semmi újdonság.

A kutatók azonban új típusú adatokat is gyűjtöttek. Megkérték a randizókat, hogy vigyenek magukkal magfónt. Aztán digitalizálták a hangfelvételeket. Így kódolni tudták a randikon használt szavakat, a nevetést, a hangszínt, a hanglejtést. Tesztelheték, hogyan jelzik a férfiak, illetve a nők, hogy érdekli őket a másik, és mivel szolgál rá a másik az érdeklődésükre.

Miről árulkodnak tehát a nyelvi adatok? Először is arról, hogyan jelzi egy férfi, illetve egy nő, hogy érdekli a másik iránt. Egy férfi esetében a vonzalom egyik jele eléggé magától értetődő: nevet a nők viccein. A másik már kevésbé magától értetődő – beszéd közben korlátozza a hangterjedelmét. Azt, hogy a monoton hanghordozást a nők sokszor férfiasnak érzékelik, kutatási eredmények is alátámasztják,

ami arra utal, hogy ha tetszik nekik egy nő, a férfiak, talán csak tudat alatt, de eltúlozzák a férfiasságukat.

A kutatók azt tapasztalták, a nők úgy jelzik az érdeklődésüket, hogy változtatják a hangmagasságukat, illetve halkabban és rövidebb intervallumokban beszélnek. Fontos nyomokat rejt a nők szóhasználata is. Egy nő aligha érdeklődik a másik iránt, ha olyan, határozatlanságot sugalló kifejezéseket használ, mint a „talán” vagy a „gondolom”.

Férfitársaim! Ha egy nő egyfolytában határozatlan mondatokkal dobálózik, bármiről legyen is szó – ha „fogjuk rá”, hogy ízlik az itala, „mondhatni”, fázik, vagy „talán” fog kérni még az előételtől –, akkor nagy összegben fogadhatunk, hogy „fogjuk rá”, „mondhatni”, „talán” egy kicsit sem jövőnk be neki.

Határozottan esélyes, hogy felkeltettük egy nő érdeklődését, ha magáról beszél. Kiderült, hogy ha egy férfi akar valamit egy nőtől, muzsika lehet a füleinek, ha azt a szót hallja a szájából, hogy „én”: ez annak jele, hogy a nő kellemesen érzi magát. Jó eséllyel az is a nő érdeklődésének jele, ha olyan közbevetésekkel él, mint „Érted?”, valamint „Úgy értem...”. Hogy miért? A kutatók megfigyelték, hogy ezek a kifejezések a másik érdeklődését hivatottak felkelteni. Ezek kedves, barátságos megnyilvánulások, és arra utalnak, hogy a másik közelebb akar kerülni hozzánk, *értik, hogy értem?*

Hogyan kommunikáljanak tehát a férfiak és a nők, ha fel akarják kelteni a partnerük érdeklődését egy randin? Az adatok szerint számos módja van annak, hogy egy férfi javítsa esélyeit. A nők az olyan férfiakat kedvelik, akik hagyják, hogy vezessék őket. Talán nem meglepő, hogy egy nő akkor számol be nagyobb eséllyel kötődés kialakulásáról, ha a férfi nevet a viccein, és az általa felhozott témákról beszélget vele, azaz nem tereli egyfolytában az őt érdeklő

dolgokra a szót.* A nők azokat a férfiakat is kedvelik, akik támogatásukról és együttérzésükről biztosítják őket. Ha egy férfi olyanokat mond, hogy „Ez óriási!”, illetve hogy „Ez nagyon király!”, lényegesen több az esélye arra, hogy a nő úgy érezze – és arról számoljon be a kutatóknak –, hogy közük van egymáshoz. Akárcsak akkor, ha olyan kifejezéseket használ, mint hogy „Az kemény!”, illetve hogy „Akkor most biztos szomorú vagy!”.

A nőknek rossz hírrel szolgálhatok, mivel az adatok megerősíteni látszanak egy visszatartó igazságot a férfiakal kapcsolatban. A társalgás minősége alig játszik szerepet abban, ahogy egy nőre reagálnak. A külső szerepe bármely más tényezőét magasan felülmúlja annak előrejelzésében, hogy egy férfi vonzalomról fog-e beszámolni a randit követően. Mindazonáltal van egy szó, amelyet a nő bizvást használhat, ha legalább egy kicsit növelni akarja az esélyeit. Ez a szó, amint azt már említettem, az „én”. A férfiak nagyobb eséllyel fognak vonzalomról beszámolni olyan nő kapcsán, aki szívesen beszél magáról. És ahogy arról már szintén esett szó, egy nő is nagyobb eséllyel fog vonzalomról beszámolni egy olyan randi után, ahol beszélt magáról. Igencsak kedvező előjel tehát, ha az első randin érdemi társalgás alakul ki a nő személyéről. Ez azt jelenti, hogy a nő kényelmesen, otthonosan érzi magát, és valószínűleg értékeli, hogy a férfi nem sajátítja ki a beszélgetést. A férfinak pedig tetszik, hogy a nő megnyílik előtte. Jó esély van a második randira.

Végezetül akadt egy egyértelműen veszjósló jel a vil-

* Van egy elméletem, amely még kidolgozás alatt áll: a Big Data a néhai Leonard Cohen minden egyes szavát megerősíti! Az unokaöccsének például ezt a tanácsot adta az udvarlással kapcsolatban: „Hallgasd őket! Aztán még többet hallgasd őket! És ha már úgy gondolsz, hogy eleget hallgattál, hallgasd őket még egy kicsit!” Ez a tanács meglehetősen emlékeztet arra, amire a kutatók jutottak.

lámrandik hanganyagának átiratában: a kérdőjel. Ha egy randin sok kérdés hangzik el, kevésbé valószínű, hogy a férfi és a nő egyaránt vonzalomról fog beszámolni. Ez kont-raintuitívnek tűnik: azt gondolhatnánk, hogy a kérdezős-ködés az érdeklődés jele. Lehet, de nem az első randin. Ott a sok kérdés többnyire az unalom jele. „Mi a hobbid?” „Hány testvéred van?” Ilyesmiket akkor kérdeznek az emberek, ha a társalgás akadozik. Egy remekül sikerült első randin viszont, valószínűleg csak a legvégén hangzik el egyetlenegy kérdés: „Találkozunk még?” Ha tényleg ez az első kérdés, akkor a válasz nagy valószínűséggel „Igen” lesz.

A férfiak és a nők nemcsak akkor beszélnek másként, ha udvarolni próbálnak egymásnak. Általában is másként beszélnek.

Egy pszichológusokból álló kutatócsoport több száz-ezer Facebook-poszt szavait elemezte. Megvizsgálták, milyen gyakorisággal használnak egy-egy szót a nők, illetve a férfiak. A végén meg tudták állapítani, hogy melyek az angol nyelv legférfiasabb, illetve legnőiesebb szavai.

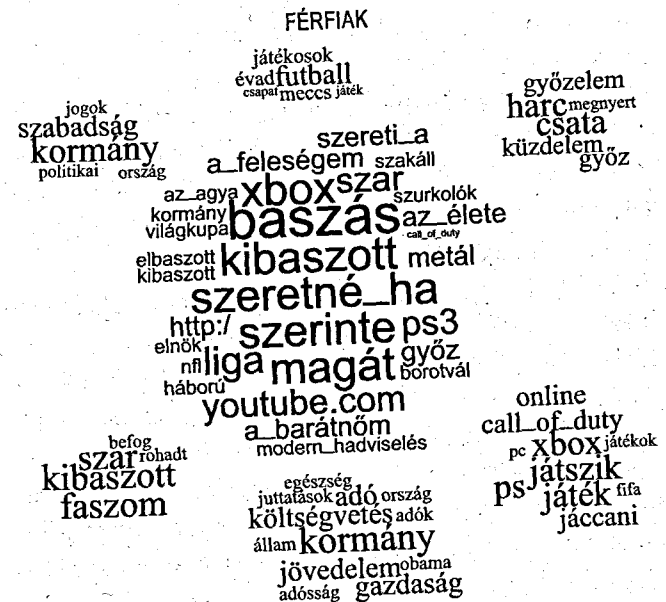
A két nem szópreferenciáinak jó része sajnos eléggé magától értetődő volt. Például a nők sokkal gyakrabban beszélnek „vásárlásról” és „a hajamról”, mint a férfiak. A férfiak sokkal gyakrabban beszélnek az „amerikai fociról” és az „Xboxról”. Valószínűleg nincs szükség egy csapat Big Datát elemző pszichológusra ahhoz, hogy ezt megtudjuk.

De azért ennél érdekesebb dolgok is kiderültek. Például az, hogy a nők sokkal többször használják a „holnap” szót, mint a férfiak, talán azért, mert a férfiaknak kevésbé megy jól az előregondolkodás. A leginkább feminin nyelv-használati jellegzetesség a „so” (úgy, annyira) „o” betűjének megnyújtása. A nők által leggyakrabban használt szavak körén belül is előkelő helyet foglal el a „soo”, a „sooo”, a „soooo”, a „sooooo”, és a „soooooo”.

Lehet, hogy csak azért, mert gyerekkoromban számos olyan nő vett körül, aki egyáltalán nem habozott ledobni olykor egy-egy „f-bombát”, de mindig is azt hittem, hogy a káromkodás az esélyegyenlőség elvének terepe. Hát nem. A férfiak által sokkal gyakrabban használt szavak között ott szerepel a „fuck” (baszni, baszás, bazmeg), a „shit” (szar, a francba, a picsába), a „fucks” (baszik), a „bullshit” (süket duma), a „fucking” (baszik, kibaszott) és a „fucker” (faszkalap).

Az alábbi két szófelhőben a férfiak, illetve a nők által leggyakrabban használt szavak szerepelnek. Minél nagyobb betűvel van szedve egy szó, annál inkább az adott nem felé billen a mérleg nyelve.

Az tetszik nekem ebben a tanulmányban, hogy az új adatok olyan mintázatok fel, amelyek régóta léteznek, de eddig nem feltétlenül tudtunk róluk. A férfiak és

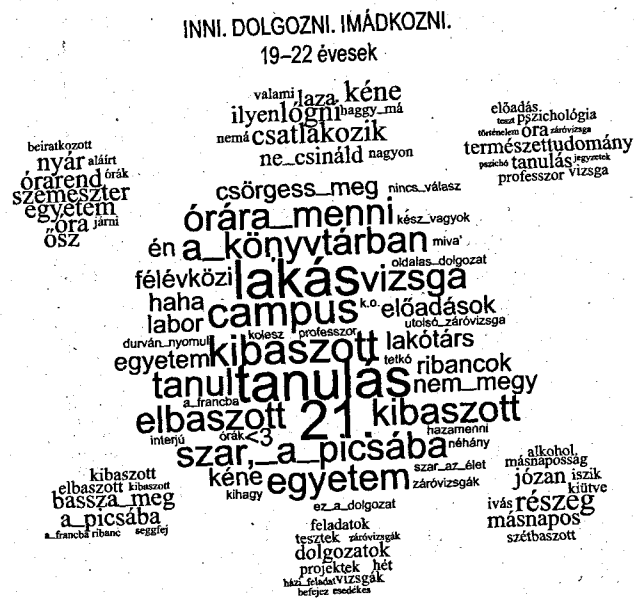




a nők mindig is másképp beszéltek, de ezek az adatok több tízezer éven át egy szempillantás alatt, a hanghullámokkal együtt eltűntek a semmiben. Ma viszont ezeket az adatokat képesek vagyunk számítógépen tárolni, és képesek vagyunk számítógéppel elemezni.

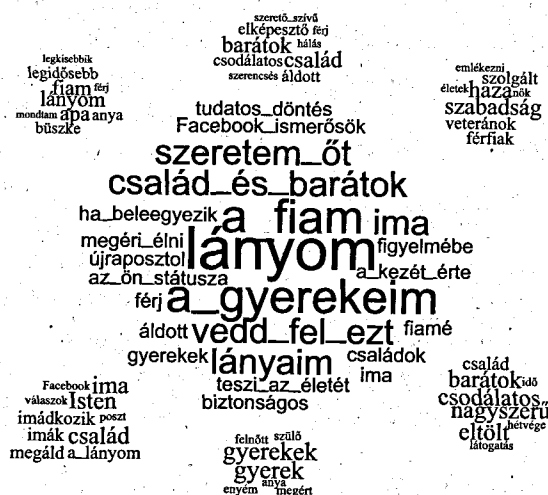
Vagy talán, tekintve, hogy férfi vagyok ezt inkább így kellett volna előadnom: „A szavak régebben kibaszottul eltűntek. De ma már megtehetjük, hogy egy időre hagyjuk a picába a focimeccset meg az Xboxot, és tanulgassuk ezt a faszágot. Ennyi, amúgy meg ki nem szarja le?”

Nem csak a férfiak és a nők beszélnek másként. Ahogy egyre idősebb lesz, ugyanaz az ember is más szavakat fog használni. A szóhasználat változása bizonyos fokig azt is jelzi, hogy miként zajlik az öregedés folyamata. Újabb szó-



INNI. DOLGOZNI. IMÁDKOZNI.
23–29 évesek





felhők következnek, még mindig ugyanabból a tanulmányból, azokkal a szavakkal, amelyeket a különböző korú emberek aránytalanul sokszor használnak a Facebookon. Úgy neveztem el ezt a grafikont, hogy „Inni, dolgozni, imádkozni.” Az emberek tizenéves korukban isznak. Huszonéves korukban dolgoznak. Harminc felett pedig imádkoznak.

A szövegelemzés egyik igen erős, új eszköze az, amit érzelelemelemzésnek hívunk. A tudósok ma már azt is meg tudják becsülni, hogy mennyire vidám vagy szomorú egy adott szövegrész:

Hogyan? Hát úgy, hogy a kutatócsoport megkér egy csomó embert, hogy pozitívként, illetve negatívként kódoljanak több tízezer angol szót. Ezzel a módszerrel derítették ki, hogy a legpozitívabb szavak közé tartozik például a „happy” (boldog, vidám), a „love” (szeret, szeretet, szerelem) és az „awesome” (klassz, eszméletlen jó). A legnegatívabbak kö-

zött ott van a „*sad*” (szomorú), a „*death*” (halál) és a „*depression*” (depresszió, válság, pangás). A kutatók ezzel létrehozták egy hatalmas szóhalmaz hangulati mutatóját.

Ennek a mutatónak a segítségével mérni tudják egy-egy szövegrész szavainak átlagos hangulatát. Ha például valaki azt írja, hogy „Boldog vagyok, szerelmes vagyok, és pompásan érzem magam”, az érzelelemzés ezt a mondatot szélsőségesen boldog szöveggént fogja kódolni. Ha valaki azt írja, hogy „Szomorú vagyok, mert az jár a fejemben, hogy hányan halnak meg világszerte, és milyen sok a depressziós ember”, az érzelelemzés ezt a mondatot szélsőségesen szomorú szöveggént fogja kódolni. Más szövegek pedig valahol a kettő között fognak elhelyezkedni a skálán.

Mennyivel leszünk tehát okosabbak attól, hogy kódoljuk a szövegek hangulatát? A Facebook adattudósai rávilágítottak egy roppant izgalmas lehetőségre: nap mint nap meg tudják becsülni egy ország GDH-ját (*Gross National Happiness*), azaz bruttó nemzeti összboldogságát. Ha az emberek állapotüzenetei jellemzően pozitívak, feltételezhetjük, hogy az ország aznap boldog. Ha inkább negatívak, feltételezhetjük, hogy az ország aznap szomorú.

A Facebook-adattudósok kutatási eredményei szerint a karácsony az év egyik legboldogabb napja. Meg kell mondanom, hogy kissé szkeptikusan viszonyultam ehhez az elemzéshez – és egy kissé szkeptikusan viszonyulok az egész projekthez. Általában véve úgy gondolom, hogy sok ember titokban szomorkodik karácsonykor, mivel magányosak, vagy a családtagjaikkal veszekednek. Még általánosabb szinten pedig eleve nem bízom a facebookos állapotfrissítésekben. Ennek okait a következő fejezetben részletesen fogom tárgyalni. Egyelőre legyen elég annyi, hogy a közösségi oldalakon hajlamosak vagyunk hazudni az életünkkel kapcsolatban.

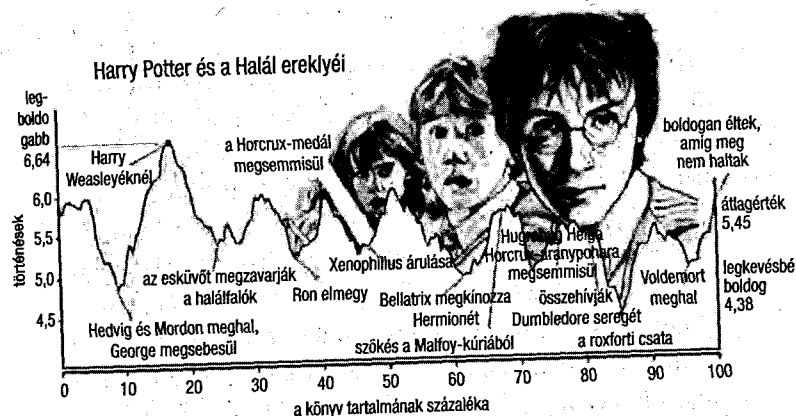
Ha ön magányos lenne, és nyomorultul érezné magát karácsonykor, talán azzal akarná nyomasztani a barátait, hogy kipoasztolja a boldogtalanságát? Élek a gyanúperrel, hogy rengetegen vannak, akiknek örömtelenül telik a karácsonyuk, de a Facebookon azt írják, hogy hálát adnak a sorsnak, amiért ilyen „csodálatosan, fantasztikusan, elképesztően boldog” az életük. Így aztán úgy fogják kódolni őket, hogy azzal jelentősen felhúzzák Amerika bruttó nemzeti összboldogságának szintjét. Ha tényleg kódolni akarjuk a GDH-t, a facebookos állapotfrissítéseken kívül jóval több adatforrást kell segítségül hívnunk.

Ezzel együtt valóban igaznak tűnik az állítás, hogy a karácsony összességében örömteli esemény. A depresszióval kapcsolatos Google-keresések, illetve Gallup-felmérések is arról tanúskodnak, hogy ez az év egyik legboldogabb napja. Az öngyilkosságok száma pedig az elterjedt városi legendát cáfolva meredeken esik az ünnepnapokon. Ha vannak olyanok, akik szomorúak és magányosak karácsonykor, sokkal több náluk a vidám ember.

Ha valaki manapság leül olvasni, legtöbbször a facebookos állapotfrissítéseket böngészi. De volt idő, nem is olyan régen, amikor az emberek történeteket olvastak, sőt még az is előfordult, hogy könyvből! Az érzelelemzésből ezen a téren is sokat megtudhatunk.

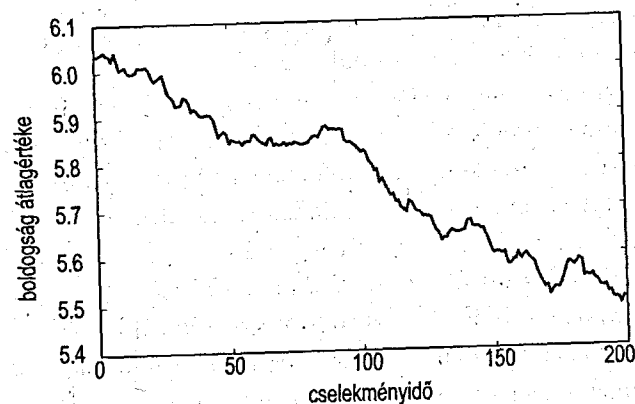
A ma a Kalifornia Egyetemen, a Berkeley School of Informationben dolgozó Andy Reagan által vezetett kutatócsoport több ezer könyv és filmforgatókönyv szövegét töltötte le. Így aztán a történetek minden egyes jelenetét kódolni tudták aszerint, hogy mennyire vidámak vagy szomorúak.

Vegyünk például a *Harry Potter és a Halál ereklyéi* című könyvet. A tudóscsoport szerint így változik a történet hangulata – a grafikonon a cselekmény legfontosabb fordulópontjai is olvashatók. Vegyük észre, hogy az érzeleme-



lemzés által kimutatott hullámhegyek és hullámvölgyek a legfontosabb eseményekhez kötődnek!

A történetek nagy többségének egyszerűbb a szerkezete. Vegyük például Shakespeare *János király* című tragédiáját! Ebben a színdarabban semmi sem alakul jól. Jánost; Anglia királyát megkérlik, hogy mondjon le a trónról. Kiközösítik az egyházból, mert nem engedelmeskedett a pápának. Kitör a háború. Meghal az unokaöccse, nem kizárt, hogy öngyilkos lett. Mások is meghalnak. Legvégül pedig Jánost megmérgezi egy elégedetlen szerzetes.



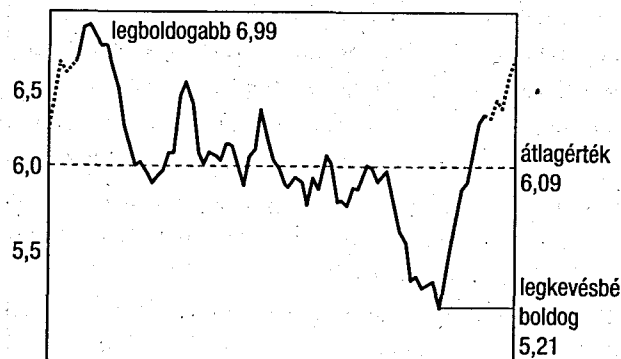
Íme, az érzelelemzés a cselekmény előrehaladásának függvényében:

Vagyis a számítógép pusztán szavak alapján képes volt kimutatni, ahogy a sztori a rossztól a rosszon át a legrosszabb felé tart.

Vagy ott van a *127 óra* című film, amelynek cselekménye nagy vonalakban ennyi:

Egy hegymászó elmegy túrázni a Utah államban található Canyonlands Nemzeti Parkba. Összehaverkodik más túrázókkal, de aztán elválnak útjaik. Emberünk egyszer csak megcsúszik, és egy kilazult kötömb a karjára zuhan. Mindent megpróbál, hogy kiszabadítsa a beszorult kezét, de hiába. Emberünk belesüpped a kétségbeesésbe. Végül amputálja a karját, és megmenekül. Megnősül, családot alapít, és a hegymászást is folytatja, de most már mindig szól valakinek, hogy hova megy.

127 óra – rendezte: Danny Boyle



Ez pedig a film érzelelemzése, szintén Andy Reagan kutatócsoportja jóvoltából:

Mit tudhatunk meg tehát sok ezernyi ilyen történet hangulatából?

A Berkeley informatikusai rájöttek, hogy a történetek túlnyomó többsége beleillik az alábbi hat, viszonylag egyszerű struktúrába. Ezek Reagan csapatának egyik kölcsönvett grafikonja alapján:

Koldusból királyfi (hangulatgörbe emelkedik).

Királyfiból koldus (hangulatgörbe süllyed).

Ember a gödörben (süllyed, majd emelkedik).

Ikarosz (emelkedik, majd süllyed).

Hamupipőke (emelkedik, majd süllyed, majd emelkedik).

Ödipusz (süllyed, majd emelkedik, majd süllyed).

Előfordulhatnak persze kisebb csavarok is, amelyeket ez az egyszerű séma nem mutat ki. A *127 óra* például a besorolása szerint „ember a gödörben”-típusú történet, még ha a történet folyamán akadnak is pillanatok, amikor a hangulat átmenetileg javul. A legtöbb történet nagyobb, általános struktúrája azonban jól illeszkedik a fenti hat kategória valamelyikébe. A *Harry Potter* és a *Halál ereklyéi* kivételt jelent ez alól.

Számos további kérdés is felmerül. Például az, hogy miként változott az idők folyamán a történetek struktúrája. Bonyolultabbak lettek-e az évek során a sztorik? Van-e különbség az egyes kultúrák között abból a szempontból, hogy milyen típusú történeteket mesélnek? Melyik típust kedvelik a legjobban az emberek? A férfiak és a nők más-más struktúrákhoz vonzódnak-e? Mi a helyzet a különböző országokban élő népekkel?

A szöveg mint adat végső soron példátlan mértékű bepillantást nyújt abba, hogy mi kell valójában a közönségnek, ami pedig különbözhet attól, amire a szerzők, kiadóvezetők, producerek szerint szükségük van. Már is van néhány olyan nyomunk, amely ebbe az irányba mutat.

Vegyük például a Wharton School (a Pennsylvániai Egyetem üzleti tudományok kara) professzorai, Jonah Berger

és Katherine L. Milkman tanulmányát arról, hogy milyen típusú történeteket osztanak meg az emberek. Azt tesztelték, hogy a pozitív vagy a negatív sztorik kerülnek-e fel jobban eséllyel a *The New York Times* e-mailen leggyakrabban elküldött cikkeinek listájára. Letöltötték egy három hónapos időszak összes *The New York Times*-cikkét, majd az érzelemelemzés módszerével kódolták a cikkek hangulatát. A pozitív sztorik között olyanok szerepeltek, mint „A tágra nyílt szemű jövények beleszeretnek a Nagy Almába”,* vagy „A filantrópia Tony-díja”. Az olyan cikkek pedig, hogy „Internetes pletykák terjednek a koreai színésznő öngyilkosságáról” vagy „Németország: meghalt a jegesmedvebébi gondozója”, nem túl meglepő módon negatív besorolást kaptak.

A két professzornak arról is volt információja, hogy hol helyezték el a lapban az adott cikket. A címlapon? A jobb felső sarokban? A bal felső sarokban? És azt is rögzítették, hogy mikor jelent meg a sztori. Kedden késő este? Hétfő reggel?

Ily módon össze tudtak hasonlítani két cikket – egy pozitívát meg egy negatívát –, amelyek nagyjából ugyanott, nagyjából ugyanakkor jelentek meg a *The New York Times* weboldalán, és megnézték, hogy az olvasók melyiket küldték el többször e-mailen.

Milyen hangulatú cikkeket osztanak meg tehát többször az emberek: pozitívakat vagy negatívakat?

Pozitív cikkeket. A szerzők ezt a konklúziót vonják le: „Minél pozitívabb töltetű a tartalom, annál nagyobb eséllyel lesz virális a sztori.”

Vegyük észre, hogy ez ellentmondani látszik annak a közkeletű újságírói vélekedésnek, mely szerint az embereket az erőszakos cselekmények és a katasztrófák vonzzák.

* Nagy Alma: New York városának beceneve. (A szerk.)

Az persze igaz lehet, hogy a tudósítások rengeteg komor hangulatú történetet zúdítanak a népre. Lehet valami a hírszerkesztőségek kedvelt mondasában is, mely szerint „Ami vérzik, azt nézik”.* A Wharton School professzorainak tanulmánya azonban arra enged következtetni, hogy az embereknek igazából több vidám sztorira lenne szükségük. Erre is lehetne fabrikálni valami bölcs mondást, mondjuk, azt, hogy „Ami mosolyog, azt megosztják” – noha ez nem rímel.

Ennyit a szomorú és a vidám szövegekről. Na de miből tudjuk, hogy mely szavak liberálisak, illetve konzervatívak, és mit árul el ez a modern hírmédiáról? Ez már egy kicsit komplikáltabb téma, amely visszavezet minket Gentzkow-hoz és Shapíróhoz. Tudják, ők voltak azok a közgazdászok, akiknek feltűnt, hogy két különböző újság már a cikk címében is másképp ír a meglegházasságról, és arra voltak kíváncsiak, hogy leleplezhető-e a szóhasználat elemzésével a politikai elfogultság.

A két becsvágyó, ifjú tudós először az amerikai Kongresszus jegyzőkönyveit vizsgálta. Mivel a beszédek átíratát már digitalizálták, le tudtak tölteni minden egyes szót, amely a 2005-ös évben elhagyta a demokrata, illetve republikánus képviselők száját. Ezután megnézhatték, hogy vannak-e olyan kifejezések, amelyeket a demokraták, illetve a republikánusok lényegesen többször használnak.

Van egypár. Íme, néhány példa:

DEMOKRATÁK ÁLTAL SOKKAL TÖBBET HASZNÁLT KIFEJEZÉSEK	REPUBLIKÁNUSOK ÁLTAL SOKKAL TÖBBET HASZNÁLT KIFEJEZÉSEK
Örökösödési adó	Haláladó

* Az angol eredetiben: „If it bleeds, it leads”. (A ford.)

DEMOKRATÁK ÁLTAL SOKKAL TÖBBET HASZNÁLT KIFEJEZÉSEK	REPUBLIKÁNUSOK ÁLTAL SOKKAL TÖBBET HASZNÁLT KIFEJEZÉSEK
Társadalombiztosítás privatizációja	Társadalombiztosítás reformja
Rosa Parks	Szaddám Huszein
Munkások jogai	Magántulajdonhoz fűződő jogok
Szegények	Kormányzati kiadások

Vajon mivel magyarázhatók a nyelvhasználat eme különbségei?

A demokraták és a republikánusok olykor ugyanazon fogalom leírására más-más kifejezést használnak. A republikánusok például 2005-ben csökkenteni akarták a szövetségi örökösödési adót, amelyet ők rendszerint „haláladóként” emlegettek (ami úgy hangzik, mintha az épp elhunytakra vetnének ki adót). A demokraták ugyanezt „örökösödési adónak” („*estate tax*”) nevezték, mert ez a szó azt sugallja, hogy a gazdagokat adóztatják. Amikor a republikánusok be akarták vezetni az egyéni nyugdíj-számlákat, a társadalombiztosítás „reformjáról” beszéltek. A demokraták számára ez a sokkal vérszjósóbban hangzó „privatizáció” volt.

A szóhasználatban mutatkozó különbségek olykor mindössze hangsúlyeltolódást jeleznek. Rosa Parks, a polgárjogi küzdelem hőse iránt például vélhetően mind a republikánusok, mind a demokraták nagy tisztelettel viseltetnek. A demokraták azonban gyakrabban beszéltek róla 2005-ben. Ugyanígy bizvást feltételezhetjük, hogy mind a republikánusok, mind a demokraták gonosz diktátornak tartják Szaddám Huszeint, a volt iraki vezetőt. De a republikánusok különösen sokat emlegették őt 2005-ben, azon igyekezetükben, hogy igazolják az iraki háború jogosságát.

Ehhez hasonlóan a Demokrata Párt számára a „munkások jogai”, illetve a „szegényekkel” való törődés kulcsfontosságú alapelveket jelentek. A republikánusok számára a „magántulajdonhoz fűződő jogok”, illetve a „kormányzati kiadások” lefaragására kulcsfontosságú alapelveket jelentenek.

És az efféle, a szóhasználatban megjelenő különbségek igen jelentősek. A Kongresszus republikánus képviselői például 2005-ben 365-ször használták a „haláladó” kifejezést, és mindössze 46 alkalommal azt, hogy „örökösödési adó”. A demokrata képviselők esetében ez az arány pont fordított: a „haláladó” kifejezést 35-ször használták, míg az „örökösödési adót” 195 alkalommal.

Gentzkow és Shapiro rájött, hogy ha a szóhasználat elárulhatja, hogy egy kongresszusi képviselő demokrata-e, vagy republikánus, akkor azt is elárulhatja, hogy egy újság balra húz-e, vagy jobbra. Ha a republikánus képviselők szívesebben használják a „haláladó” kifejezést, mert arról akarják meggyőzni az embereket, hogy ellenezzék ezt az adónemet, akkor lehet, hogy ugyanez a helyzet a konzervatív lapokkal is. A viszonylag liberális *The Washington Post* 13,7-szer gyakrabban ír „örökösödési adóról”, mint „haláladóról”. A konzervatív *Washington Times* nagyjából egyforma gyakorisággal használja a két kifejezést.

Hála az internet csodáinak, Gentzkow és Shapiro az ország rengeteg újságjának nyelvhasználatát elemezhetette. Ebben két weboldal, a newslibrary.com és a proquest.com+ volt a segítségükre, amelyek összesen 433 újság anyagát digitalizálták. Összesen ezer, politikai töltettel bíró kifejezés előfordulásait számolták össze: így mérték a lapok politikai elfogultságát. A mérés alapján a *Philadelphia Daily News* bizonyult a legliberálisabb újságnak; a legkonzervatívabbnak pedig a montanai *Billings Gazette*.

Most, hogy megvan a média elfogultságának első átfogó mérése, amely a sajtóorgánumok széles skáláját öleli fel, ráterhetünk a sajtóval kapcsolatos legfontosabb kérdés megválaszolására: miért húznak egyes kiadványok balra, mások pedig jobbra?

A két ifjú közgazdász hamar becserkészte a mindent eldöntő, kulcsfontosságú tényezőt: az adott terület politikai irányultságát. Kiderült, hogy ha egy terület általában véve liberálisabb, mint például Philadelphia és Detroit, akkor jellemzően a terület vezető lapja is liberális. Ha egy terület általában konzervatívabb, mint például Billings vagy a texasi Amarillo, akkor jellemzően a terület vezető lapja is konzervatív. Az adatok tehát erősen arra utalnak, hogy az újságok hajlamosak ímegadni olvasóiknak azt, amire szükségük van.

Talán azt gondolják, hogy a laptulajdonos bír némi befolyással a tudósítások irányultságára, pedig főszabály szerint a tulajdonos személye valójában jóval kisebb hatással van egy újság politikai elfogultságára, mint gondolnánk.

Figyeljük meg, mi történik, ha ugyanaz a személy vagy cég különböző piacokon is rendelkezik lapokkal! Vegyük például a The New York Times Companyt! Övék a Gentzkow és Shapiro elemzése szerint liberális elfogultságú *The New York Times*, amely abban a New York városban jelenik meg, ahol a lakosság durván 70 százaléka liberális. A tanulmány írásakor ugyancsak az övék volt a Gentzkow és Shapiro elemzése szerint konzervatív elfogultságú *Spartanburg Herald-Journal* is, amely abban a dél-karolinai Spartanburgban jelenik meg, ahol a lakosság durván 70 százaléka republikánus. Persze vannak kivételek is. Rupert Murdoch cége, a News Corporation tulajdonában van például a *New York Post*, amelyet nagyjából mindenki konzervatívnak minősítene. A kutatási eredmények azonban összességében

mégis azt sejtetik, hogy egy újság irányultságát sokkal nagyobb mértékben határozza meg a piac, mint a tulajdonos.

A tanulmány alapjaiban változtatta meg mindazt, amit a hírmédiáról gondolunk. Sokan, különösen a marxisták, hagyományosan úgy vélekednek, hogy az amerikai sajtót gazdag emberek és nagyvállalatok tartják a markukban a tömegek befolyásolása céljából, vagyis azért, hogy saját politikai nézeteik felé tereljék az embereket. Gentzkow és Shapiro tanulmánya viszont azt sugallja, hogy a tulajdonosokat elsősorban nem ez motiválja. Az amerikai sajtó inkább azt adja meg a tömegeknek, amire szükségük van, hogy a tulajdonosok még jobban meggazdagodhassanak.

Ja, és van itt még egy kérdés – egy nagy vitákat kiváltó és talán az eddigieknél is provokatívabb kérdés. Összességében inkább balra, vagy inkább jobbra húz-e az amerikai hírmédia? Inkább liberális a média, vagy inkább konzervatív?

Gentzkow és Shapiro úgy találta, hogy az újságok balra húznak. Vagyis az átlagos újság szóhasználata jobban hasonlít egy demokrata képviselőéhez, mint egy republikánuséhoz.

„Aha! Én megmondtam!” – kiálthatnak fel most a konzervatív olvasók. Közülük sokan már régóta gyanítják, hogy a sajtó elfogult, és manipulálni próbálja a tömegeket, hogy tegyék magukévá az ő baloldali álláspontjukat.

Nem, nem éppen ez a helyzet – mondják a tanulmány szerzői. Hanem az, hogy a liberális elfogultság az újságolvasó közönség igényeihez van kalibrálva. Az újságolvasó közönség pedig összességében egy kissé balra húz. (Erre adataik is vannak.) Ezért összességében kissé az újságok is balra húznak, igazodva az olvasók által követelt állásponthoz.

Nincs semmiféle nagy összeesküvés. Egyszerűen csak kapitalizmus van.

Gentzkow és Shapiro eredményei azt sugallják, hogy a hírmédia sokszor úgy működik, mint bármely más üzletág ezen a bolygón. Akár a supermarketek, amelyek kifürkészik, hogy milyen jégkrémre vágyik a nép, és azzal pakolják tele a polcokat, az újságok is kifürkészik, hogy milyen nézetekre vágyik a nép, és azokkal pakolják tele az oldalait. „Ez egyszerűen üzlet” – mondta nekem Shapiro. Hát ilyesmiket tudhat meg az ember, ha a mégoly összetett, nyakatekert híreket, elemzéseket, véleménycikkeket számszerűsíti, és alkotóelemeikre bontja: vagysis szavakra.

A KÉP MINT ADAT

Korábban a kutatók, illetve az üzletemberek körében a közvélemény-kutatás volt az adatgyűjtés bevett módja. A számadatok, kérdőíveken bejelölt rubrikákból származó információk szépen formába öntve érkeztek hozzájuk. Ma azonban már más a helyzet. Az egyszerű, könnyen átlátható, strukturált, felmérésen alapuló adatok kora lejárt. Ebben az új korszakban azok a rendezetlen, keszekusza nyomok váltak elsődleges adatforrássá, amelyeket lépten-nyomon hátrahagyunk, miközben éljük az életünket.

Amint azt már láttuk, a szó adat. A kattintás adat. A link adat. Az elgépelés adat. Az álmainkban megjelenő banán adat. A hangszín adat. A zihálás adat. A szívverés adat. A lép mérete adat. Amellett fogok érvelni, hogy mind közül az internetes keresés a legeslégárulkodóbb adat.

És mint kiderült, a kép is adat.

Ma már nemcsak a polcon porosodó könyvekbe, folyóiratokba zárt szavakat, hanem a képeket is digitalizálják, azaz kiszabadítják azokat az albumok, kartondobozok fogásából. Ezeket is bitekké alakítják, és feltöltik a felhőbe.

Így nemcsak a szövegek tarthatnak nekünk történelemórát – megvilágítva például, hogy miképp változott az emberek nyelvhasználata –, hanem a képek is – például úgy, hogy nyomon követik, hogyan változott az, ahogy az emberek az idők folyamán a fényképezésnek pózoltak.

Vegyünk egy, a Brown Egyetem és a Berkeley Egyetem négy informatikusa által készített, igen ötletes tanulmányt! A szerzőknek kapóra jött a digitális kor egyik leleménye: számos középiskola beszkenelte és online hozzáférhetővé tette az évkönyveit. A kutatók 949 évkönyvet találtak a neten az 1905 és 2013 közötti időszakból. Ezek az évkönyvek több tízezer végzős diák tablóképét tartalmazták. A kutatócsoport egy szoftver segítségével megalkotta minden egyes évtized „átlagos” arcát. Más-ként fogalmazva: meghatározták a lefényképezett diákok orrának, szemének, ajkának és hajának átlagos elhelyezkedését és formáját. Íme, az elmúlt száz év átlagos arcai, nemek szerinti bontásban



Feltűnik valami? Az amerikaiak – és különösen a nők – elkezdtek mosolyogni. A huszadik század elején még szinte megkövült arccal meredtek a fotósra, de az évszázad végén már fülig ért a szájuk.

Vajon mi lehet a változás oka? Talán boldogabbak lettek az amerikaiak?

Dehogy. A kérdés megválaszolásához más tudósok segítettek hozzá bennünket. Az ok, legalábbis szerintem, egészen lenyűgöző. Arról van szó, hogy amikor a fényképezőgép még új találmány volt, az emberek úgy tekintettek a fotóra, mint egy festményre. Nem is tudták máshoz hasonlítani. Így a fényképek modelljei a festmények modelljeit utánozták. És mivel a festményekhez modellt ülők nemigen mosolyoghattak egyfolytában a hosszú órákon át, amíg a kép készült, inkább komoly képet vágtak. Így hát a lefotózott emberek is komolyan néztek a kamerába.

Mitől változott ez meg végül? Természetesen az üzlet, a profit és a marketing miatt. A huszadik század közepén a filmet és fényképezőgépeket is gyártó Kodaknak elege lett abból, hogy az emberek csak igen korlátozott számú fotót készítenek, és kidolgozott egy stratégiát arra, hogy ezen változtasson. A Kodak reklámjaiban elkezdtek a boldogság képzetét társítani a fotózáshoz. Rá akarták szoktatni az embereket, hogy másoknak is meg akarják mutatni, milyen jól érezték magukat. A középiskolai évkönyvek sok-sok mosolygós fotója ennek a sikeres hirdetési kampánynak az eredménye (ahogy a mai Facebook- és Instagram-fotók zöme is).

De a fényképből mint adatból jóval többet is megtudhatunk annál, hogy mikortól mondják az érettségizők, hogy „Csíííí!”. Meglepő módon akár a gazdasági helyzetről is képet alkothatunk a képek alapján.

Volt egyszer egy tudományos dolgozat, amely azt a provokatívba hajló címet kapta, hogy „A gazdasági növekedés mérése a világürből”. Ha egy tanulmánynak ilyen címe van, mérget vehetnek rá, hogy én elolvasom! A szerzők – J. Vernon Henderson, Adam Storeygard és David N. Weil – azzal az észrevétellel kezdik, hogy sok fejlődő or-

szágban nem túl eredményes a bruttó nemzeti össztermék (GDP) mérése. Ennek az az oka, hogy a gazdasági tevékenység igen jelentős hányada feketén zajlik, a mérést végző kormányzervek erőforrásai pedig korlátozottak.

A szerzőknek támadt egy meglehetősen rendhagyó ötletük. Kitalálták, hogy az éjszakai fényerősség alapján fogják mérni a GDP-t. Az információt az amerikai légierő egyik műholdja szolgáltatta, amely naponta 14-szer kerül meg a Földet.

Mi köze az éjszakai fénynek a GDP-hez? Hát az, hogy a világ legszegényebb részein élők számára nagy gondot jelent a villanyszámla fizetése. Ebből pedig az következik, hogy ha romlik a gazdasági helyzet, akkor a háztartások, illetve a falvak éjszakánként drasztikusan kevesebb fényerősséget engedélyeznek maguknak.

Indonéziában, az 1998-as ázsiai pénzügyi válság idején nagyot zuhant az éjszakai fényerősség. Dél-Koreában 1992 és 2008 között 72 százalékkal nőtt, ami egybevág ennek az időszaknak a figyelemre méltó gazdasági teljesítményével. Észak-Koreában ugyanebben az időszakban csökkent a fényerősség, ami egybevág ennek az időszaknak a gyászos gazdasági teljesítményével.

1998-ban Madagaszkár déli részén óriási rubin- és zafírelőhelyet fedeztek fel. Ennek köszönhetően Ilakaka városa, amely addig alig volt több egy kamionpihenőnél, komoly kereskedelmi központtá nőtte ki magát. Ilakaka 1998 előtt gyakorlatilag koromsötét volt éjszakánként, a következő öt év alatt viszont ugrásszerűen megnőtt a fényerősség.

A tanulmány szerzői maguk is belátják, hogy ez közel sem hibátlan módszer a gazdasági teljesítmény mérésére. Nyilvánvaló, hogy lehetetlen pontosan felmérni a gazdasági helyzetet pusztán az alapján, hogy mennyi fényt észlelnek a műholdak éjszakánként. A szerzők nem is javasolják,

hogy fejlett országokra, így például az Egyesült Államokra alkalmazzuk a módszerüket, ahol a létező gazdasági adatok sokkal megbízhatóbbak. És igazság szerint arra jutottak, hogy a harmadik világbeli országok tekintetében is csak olyan mértékben tudják hasznát venni az éjszakai fényerősségmérésnek, mint a hivatalos adatoknak. Ha viszont a hibás kormányzati adatokat kombináljuk a tökéletlen fényerősségadatokkal, jobb közelítéssel becsülhetjük meg a gazdasági teljesítményt, mint ha bármelyik forrást önmagában vennénk alapul. Vagyis a világűrből lőtt képek segítségével többet tudhatunk meg a fejlődő országok gazdaságáról, mint eddig.

Joseph Reisingert, az informatika halk szavú doktorát a fenti tanulmány szerzőihez hasonlóan bosszantotta, hogy a fejlődő országok gazdaságára vonatkozó adathalmazok ennyire hiányosak. Említ egy példát, amikor is Nigéria 2014 áprilisában frissítette GDP-becslését, és ennek során a gazdaság olyan szektorait is figyelembe vették, amelyek a korábbi becslésekből kimaradtak. A becsült GDP így 90 százalékkal magasabb lett.

„Afrika legnagyobb gazdaságáról van szó” – mondta lassan emelkedő hangerővel Reisinger –, „de a legalapvetőbb dolgot sem tudjuk róluk!”

Találni akart egy módszert arra, hogy pontosabb képet kaphasson Nigéria gazdasági teljesítményéről. A megoldása remek példája annak, hogyan lehet újragondolni, mi is minősül adatnak, ahogy azt is jól példázza, milyen hatalmas haszon rejlik az efféle újragondolásban.

Reisinger alapított egy Premise nevű céget, amely okostelefonnal felfegyverzett emberek sokaşágát alkalmazza a fejlődő országokban. Az a dolguk, hogy lefotózzanak mindent, ami gazdasági szempontból érdekes lehet.

Fényképeznek például benzinkutakat vagy a szupermarketek gyümölcsös ládáit. Újabb és újabb fényképeket készítenek ugyanarról a helyszínről. A képeket elküldik a Premise-nek, ahol az alkalmazottak egy másik csoportja – az informatikusok – adatokká alakítják a felvételeket. A Premise elemzőgárdája bármit képes kódolni: hogy milyen hosszú a sor a benzinkutakon, hogy hány almát raktak ki a szupermarketben, hogy mennyire érett az alma, vagy hogy mennyibe kerül. A Premise a különféle tevékenységekről készült fotók alapján képes becsléseket készíteni a gazdasági teljesítményre, illetve az inflációra vonatkozóan. A fejlődő országokban a benzinkúton kígyózó sor azt jelzi, hogy baj van a gazdasággal. Erről tanúskodik az is, ha a boltban nincs alma, vagy van, de éretlen. A terepen készült fotóknak köszönhetően a Premise már jóval a hivatalos adatok közzététele előtt tudott a 2011-es kínai élelmiszer-inflációról, majd az azt követő, 2012-es deflációról is.

A Premise bankoknak és hedge fundoknak* adja el információit, és a Világbankkal is együttműködik.

Mint sok más jó ötlet, a Premise is olyan ajándék, amelynek folyamatosan újabb és újabb hozadéka van. A Világbankot például nemrégiben érdekelni kezdte a fülöp-szigeteki illegális cigarettagyártás nagyságrendje. Konkrétan arra voltak kíváncsiak, hogy milyen hatást ért el a kormány akciója, amelynek részét képezték a váratlan rajtaütések is: azokra a gyártókra próbáltak lecsapni, amelyek kibújtak a jövedéki adó fizetése alól. A Premise-nek támadt egy furfangos ötlete: lefotóztatták embereikkel az utcán látott cigarettás dobozokat, és megnézték, van-e rajtuk zárjegy.

* Hedge fund: alacsony szabályozottságú, zárt körű befektetési alap. (A szerk.)

Arra jutottak, hogy a feketegazdaságnak ez az ága 2016-ra igencsak megcsappant az előző évhez képest. A kormány erőfeszítéseit tehát siker koronázta, de mivel földalatti tevékenységről – illegális cigarettáról – van szó, új adatokra volt szükség ahhoz, hogy ez kiderüljön.

Amint azt már láttuk, digitális korunkat annak bőszerű gondolkodása jellemzi, hogy mi minősül adatnak, és ezekben az új típusú információkban sok jelentős felismerés rejlik. Annak kiderítése, hogy mi áll a média elfogultsága mögött, mitől sikerül jól az első randi vagy mi a valós helyzet a fejlődő országok gazdaságában, még csak a kezdet.

Nem véletlen, hogy az új adatok a felismerések mellett egy rakás pénzt is fialtak, kezdve Brin és Page urak tízmilliárdjával. Joseph Reisingernek sem megy rosszul. A megfigyelők számítása szerint a Premise évente több tízmillió dollárnyi bevételt termel, és a befektetők nemrégiben 50 millió dollárt lapátoltak a cégbe. Ez azt jelenti, hogy egyes befektetők a világ legértékesebb olyan cégei közé sorolják a Premise-t, amelyek főként fotózással és fotóeladással foglalkoznak, vagyis ugyanabban a ligában, amelyben a *Playboy* is indul.

Más szóval: mind a tudósok, mind az üzletemberek szemében hatalmas értéket képvisel a ma már hozzáférhető, újszerű adattípusok felhasználása, illetve az adat fogalmának kitágítása. Egy adattudós ma már nem korlátozhatja tevékenységét az adatok szűkebben, illetve hagyományos értelemben vett körére. Manapság a szupermarketben kigyózó sorról készült fotó értékes adat. Adat a gyümölcsös ládák tartalma. Adat az alma érettségének szintje. Adatok a világúrból készített fotók. Adat a száj görbülete. Adat az egész világ.

És mindezen adatoknak köszönhetően végre-valahára átlátunk az emberek hazugságain.

4. fejezet

Digitális igazságszérum

Mindenki hazudik.

Az emberek hazudnak arról, hogy mennyit ittak hazafelé menet. Hazudnak arról, hogy hányszor járnak le a konditerembe, hogy mennyibe került az új cipőjük, hogy olvasták-e azt a bizonyos könyvet. Beteget jelentenek, holott kutyájuk bajuk. Azt mondják, majd hívnak, aztán nem hívnak. Azt mondják, nem veled van a baj, pedig veled van a baj. Azt mondják, szeretnek, pedig nem. Azt mondják, boldogok, pedig nagyon el vannak kenődve. Azt mondják, a nőket szeretik, pedig valójában a férfiakat szeretik.

Az emberek hazudnak a barátaiknak. Hazudnak a főnököknek. Hazudnak a gyerekeiknek. Hazudnak a szüleiknek. Hazudnak az orvosuknak. Hazudnak a férjüknek. Hazudnak a feleségüknek. Hazudnak önmaguknak.

Még szép, hogy a közvélemény-kutatóknak is hazudnak. Itt van önnek egy kis villámfelmérés:

Puskázott-e már valaha vizsgán?

Ábrándozott-e már valaha arról, hogy megöl valakit?

.....

Megkísértette, hogy hazudjon? A felmérések válaszadói közül sokan hallgatják el vagy kisebbitik a kínos szokásokat és gondolatokat. Jó benyomást akarnak kelteni, még

akkor is, ha névtelenül szolgáltatnak adatokat. A jelenség neve „társadalmi elvárások által generált torzítás”.

Már 1950-ben készült egy fontos tanulmány, amely igen erős bizonyítékokkal szolgált arra vonatkozóan, miként eshetnek áldozatul a felmérések ennek a torzításnak. A kutatók utánanézték a hivatalos adatforrásokban, hogy Denver lakóinak hány százaléka szokott szavazni, hány százalék adakozik karitatív célokra, és hány százaléknak van könyvtári olvasójegye. Ezután felmérést végeztek a lakosság körében, hogy megállapíthassák, egybevág-e a válaszok a hivatalos adatokkal. Az eredmény akkor még sokkolta a kutatókat. Egészen mást mutatott, mint a város lakóiról összeszedett adatok. A kérdőívet névtelenül töltötték ki, mégis rengetegen túloztak a választási regisztrációjukat, választói viselkedésüket, illetve a jótékonykodásukat illetően.

	A FELMÉRÉS ADATAI SZERINT	A HIVATALOS ADATOK SZERINT
Regisztrált szavazó	83%	69%
Szavazott a legutóbbi elnökválasztáson	73%	61%
Szavazott a legutóbbi polgármester-választáson	63%	36%
Van könyvtári olvasójegye	20%	13%
Adakozott a Community Chest nevű jótékony szervezet legutóbbi gyűjtésén	67%	33%

Változott-e bármi is az azóta eltelt 65 évben? Talán anyni, hogy ma, az internet korában már nem kínos, ha az ember nincs beiratkozva a könyvtárba. De ha az változott is, hogy mi minősül kínosnak, illetve kíváncsnak, az em-

berek továbbra is előszeretettel verik át a közvélemény-kutatókat.

Nemrégiben, a Marylandi Egyetemen végzett felmérés keretében számos kérdést feltettek az ott végzetteknek az egyetemmel kapcsolatos tapasztalataikról. A válaszaikat összevetették a hivatalos adatokkal. Az öregdiákok következetesen hamis információkat szolgáltattak annak érdekében, hogy jobb fényben tüntessék fel magukat. Alig 2 százalékuk ismerte be például, hogy közepesnél rosszabb átlaggal diplomázott. (Valójában 11 százalék ért el ilyen eredményt.) Azt pedig, hogy az előző évben adományt küldött az egyetemnek, 44 százalékuk állította. (Valójában ezt mintegy 28 százalékuk tette meg.)

És a hazugság nyilván szerepet játszhatott abban is, hogy a közvélemény-kutatások nem jelezték előre Donald Trump 2016-os győzelmét. A közvélemény-kutatások átlagosan 2 százalékponttal becsülték alá a támogatottságát. A megkérdezettek egy része talán kínosnak érezte volna, ha bevallja, hogy rá fog szavazni. Egyesek talán bizonytalannak mondták magukat, holott mindvégig Trumphoz húztak.

Vajon miért informálják félre az emberek névtelenül is a közvélemény-kutatókat? Megkérdeztem erről Roger Tourangeau-t, a Michigani Egyetem professzor emeritusát, a társadalmi elvárások által generált torzítás talán leghíresebb szakértőjét. A professzor elmagyarázta, hogy a problémának fontos eleme a „kegyes hazugságra” való hajlamunk. „A való életben az emberek körülbelül minden harmadik szava hazugság. Márpedig a szokás nagy úr, így a közvélemény-kutatók sem ússzák meg” – mondja.

Aztán van még egy további furcsa szokásunk is: az, hogy időnként magunknak is hazudunk. „Az ember nem szívesen vallja be magának, hogy egyetemista korában elfuserált alak volt” – mondja erről Tourangeau.

Az önbecsapás lehet a magyarázata annak is, hogy miért állítja magáról olyan sok ember, hogy átlagon felüli. Hogy mekkora gond ez? Egy cégnél a mérnökök több mint 40 százaléka azt állította, hogy a legfelső 5 százalékba tartozik. Az egyetemi professzorok több mint 90 százaléka vélekedik úgy, hogy átlagon felüli munkát végez. A középiskolák végzőseinek egynegyede a legfelső 1 százalékba sorolja magát abból a szempontból, hogy mennyire tud bánni az emberekkel. És aki áltatja magát, az a felmérésekben sem tud őszinte lenni.

Egy további tényező is közrejátszik a közvélemény-kutatóknak adott hazug válaszainkban: az az igen erős vágy, hogy jó benyomást tegyünk egy idegenre, esetünkben a kérdezőbiztosra, már ha van ilyen. Tourangeau így fogalmaz: „Képzeld el, hogy belép egy hölgy, aki úgy néz ki, mint a kedvenc nagynénje! Tényleg el akarja mondani a kedvenc nénikéjének, hogy az előző hónapban füvet szívtott?”* És ön szívesen bevallaná, hogy sajnálta a pénzt a jó öreg alma materétől?

Mindezek miatt minél személytelenebbek a körülmények, annál őszintébbek a válaszok. Az internetes felmérések több igaz szót csalogatnak elő az alanyokból, mint

* A hazudozás másik oka az lehet, hogy a megkérdezettek egyszerűen szórakoznak a felméréssel. A tinédzsereknél ez különösen nagy gondot okoz, és még komplikáltabbá teszi e korosztály megértését. A kutatók például annak idején kimutatták, hogy ha valaki örökbe fogadott gyerek, az korrelál bizonyos negatív tinédzserkori viselkedésművelettel (drogozás, alkoholfogyasztás, iskolakerülés). A későbbi kutatások során viszont kiderült, hogy a korreláció teljes mértékben magyarázható a minta 19 százaléka kitévő, önmagát örökbe fogadottként kategorizáló tinédzserrel, akiket valójában nem fogadtak örökbe. Az ezt követő újabb kutatások kimutatták, hogy a tinédzserek elég jelentős hányada vallja magát 210 centinél magasabbnak, 180 kilónál nehezebbnek vagy éppen háromgyerekes szülőnek. Egy felmérés során kiderült, hogy az állításuk szerint művégtaggal élő egyetemisták 99 százaléka csak viccelt.

a telefonosak, amelyek viszont még mindig jobbak, mint a személyes kérdezés. Az emberek több mindent vallanak be, amikor egyedül vannak, mint amikor nem.

Am ha érzékeny témáról van szó, akkor az általunk használt módszertől függetlenül is jelentős mennyiségű hamis információra számíthatunk. Tourangeau professzor itt olyan szót használt, amellyel a közgazdászok elég sűrűn dobálóznak: „ösztönző”. Az embereket semmi sem ösztönzi arra, hogy igazat mondjanak a felméréseken.

Hogyan tudhatjuk meg tehát, mit gondolnak és cselekszenek valójában embertársaink?

Előfordul, hogy vannak olyan hivatalos adatforrások, amelyekben utána tudunk nézni az igazságnak. A jótekonkodásról például hiába hazudnak az emberek, hiszen maguktól a karitatív szervezetektől is beszerezhetjük a valós számokat. Ha viszont olyan viselkedésműveletéről akarunk tájékozódni, amelyek nem szerepelnek semmilyen nyilvántartásban, vagy arra vagyunk kíváncsiak, hogy mit gondolnak, hisznek, éreznek, mire vágnak az emberek valójában, akkor nincs más információforrás, mint az, amit méltóztatnak elmondani a közvélemény-kutatóknak. Legalábbis eddig így volt.

Ez a Big Data második komoly képessége: bizonyos online adatforrásokkal olyasmiket közölnek az emberek, amit sehol máshol nem vallanának be. Ezek az adatforrások afféle digitális igazságszérumként működnek. Gondoljunk csak a Google-keresésekre! Megvan még, hogy milyen körülmények között nyílnak meg a legjobban az emberek? Online? Pipa. Egyedül? Pipa. Kérdezőbiztos jelenléte nélkül? Pipa.

És van a Google-kereséseknek egy további előnyük is, amely az igazság felé tereli a felhasználókat: az ösztönzők. Ha valakinek bejönnek is a rasszista viccek, semmilyen

ösztönző sem hat abba az irányba, hogy ezt a politikailag inkorrekt tényállást megossza a közvélemény-kutatókkal. Arra azonban ösztönözve van, hogy megkeresse a neten a legjobb rasszista vicceket. Ha valaki depressziósnak érzi magát, semmilyen ösztönző sem hat abba az irányba, hogy ezt bevallja egy felmérésen. Arra azonban ösztönözve van, hogy kifaggassa a Google-t a tünetekről és a lehetséges gyógymódokról.

Még ha hazudunk is önmagunknak, a Google tudhatja az igazságot. Például ön és néhány szomszédja nyugodtan gondolhatja úgy pár nappal a választás előtt, hogy majd elautóznak a szavazóhelyiséghez, és leadják a voksukat. De ha nem kerestek rá a neten olyan információkra, hogy hol és hogyan kell szavazni, akkor a magamfajta adattudósok ebből kikövetkeztethetik, hogy az önök környékén várhatóan alacsony lesz a részvétel. Ugyanígy előfordulhat, hogy ön még nem vallotta be magának, hogy depressziós tünetei vannak, miközben már rákeresett a sírásrohamokra és arra, hogy miért esik nehezeire felkelni az ágyból. Ennek ellenére meg fog jelenni egy adott terület depresszióval kapcsolatos kereséseiben, amelyeket már elemeztem ebben a könyvben.

Gondoljon csak ki-ki a saját Google-élményeire! Arra tippelek, hogy olykor ön is begépelte már olyasmit a keresősorba, amely olyan szalonképtelen szokásokról és gondolatokról árulkodik, amilyeneket decens társaságban habozna bevallani. Igazság szerint az adatok perdöntő bizonyítékkal szolgálnak arra, hogy az amerikaiak túlnyomó többsége igencsak személyes dolgokat árul el a Google-nak. Többször keresnek például a neten pornót, mint ahányszor megnézik az időjárás-jelentést. Amit egyébként nem könnyű összehámozni a felmérések adataival, hiszen a férfiaknak csak mintegy 25, a nőknek pedig mintegy 8 százaléka ismeri be, hogy szokott pornót nézni.

Akkor is feltűnhetett már önnek, hogy a Google-t bizonyos fokú őszinteség jellemzi, amikor a keresőmotor automatikusan ki akarta egészíteni az ön által megkezdett kifejezést. Ezek a javaslatok a mások által leggyakrabban beírt kereséseken alapulnak. Vagyis az automatikus kiegészítés jelzi, hogy mire szoktak rákeresni az emberek. Ez persze kissé félrevezető tud lenni, hiszen a Google nem javasol bizonyos, illetlennek ítélt szavakat, mint például olyasmiket, hogy „fasz”, baszás” vagy „pornó”. Az automatikus kiegészítés tehát azt sugallja, hogy az emberek Google-s gondolatvilága nem is olyan pikáns és sikamlós, mint valójában. Ám még így is sokszor felszínre törnek bizonyos érzékeny témák.

Ha például beírjuk, hogy „miért...”, az első két automatikus kiegészítés ma az, hogy „miért kék az ég?”, és „miért vannak szökőévek?”, ami arra utal, hogy ez a két leggyakoribb folytatás. A harmadik pedig ez: „miért kakálok zöldet?”. A kereséseink automatikus kiegészítése olykor egészen felkavaró tud lenni. Ha például valaki ma bepötyögi azt, hogy „Normális, ha meg akarom...”, akkor az elsőként javasolt folytatás az, hogy „ölni”. Az után pedig, hogy „Normális, ha meg akarom ölni...”, az az első javaslat, hogy „a családomat”.

Ön szerint ez még nem elegendő bizonyíték arra, hogy a Google-keresések más képet festenek a világról, mint amelyet általában a saját szemünkkel látunk? Akkor időzőn el egy kicsit az azzal kapcsolatos kereséseknél, hogy valaki megbánta-e a gyerekvállalással kapcsolatos döntését! A döntés előtt állók egy része fél, hogy rosszul választ. A kérdés szinte mindig az, hogy nem fogják-e megbánni, ha *nem lesz* gyerekük. Ezt hétszer akkora valószínűséggel kérdezik meg a Google-től az emberek, mint azt, hogy nem fogják-e megbánni, ha *lesz* gyerekük.

Akik túl vannak a nagy elhatározáson – hogy szülessen-e gyermekük, illetve fogadjanak-e örökbe –, azok időnként bevallják a Google-nak, hogy már szánják-bánják a döntésüket. Lehet, hogy ez egy kicsit sokkolóan hangzik, de a döntés után a visszajára fordul a statisztika. Azok a felnőttek, akiknek van gyerekük, 3,6-szor nagyobb eséllyel mondják a Google-nak, hogy megbánták a döntésüket, mint a gyermektelen felnőttek.

Egy kitételt érdemes a fejezet végéig észben tartaniuk: a Google időnként elfogult a nem helyénvaló gondolatok javára, amelyeket az emberek aligha tudnának megbeszélni mással. Ám ezzel együtt is igaz, hogy ha megpróbáljuk leleplezni a rejtett gondolatokat, hasznunkra válhat a Google-nak az a képessége, hogy képes előtérni ezeket. És a jelentős arányeltolódás, amelyre a gyerekvállalás, illetve a gyermektelenség megbánásának kapcsán fény derült, azt sejteti, hogy a nem helyénvaló gondolatoknak ez esetben komoly súlyuk van.

Álljunk meg itt egy pillanatra, és gondoljuk végig, mit is jelent az egyáltalán, hogy valaki ezt írja a keresőbe: „megbántam, hogy gyereket vállaltam”. A Google olyan adatforrásként kínálja magát, ahonnan közvetlen információhoz juthatunk a várható időjárásról, hogy ki nyerte a tegnapi meccset, vagy hogy mikor épült a Szabadság-szobor. De az is előfordul, hogy anélkül pötyögjük be cenzúrázatlan gondolatainkat, hogy sok reményt fűznénk ahhoz, hogy a Google segíthet. A keresősor ilyenkor a gyóntatószék szerepét tölti be.

Évente sok ezren keresnek rá például olyasmikre, hogy „utálom a hideget”, „idegesítenek az emberek” vagy éppen hogy „szomorú vagyok”. Persze ez a sok ezer keresés arra, hogy „szomorú vagyok”, mindössze apró töredékét jelenti meg a sok százmillió embernek, akiket az év folyamán

elfog a szomorúság. Kutatásaim szerint a nem tájékozódó célzatú, hanem gondolatokat kifejező keresések csak nagyon csekély hányadát fedik le azoknak, akiknek eszükbe jut ez a gondolat. A kutatásaim azt is sejtetni engedik, hogy az évi 700 ezer alkalom, amikor az amerikaiak rákeresnek arra, hogy „megbántam, hogy gyereket vállaltam”, csak nagyon csekély hányadát fedi le azoknak, akiknek eszükbe jut ez a gondolat.

Nyilvánvaló, hogy szülőnek lenni sok ember számára, talán a többség számára hatalmas öröm. És hiába fél az anyukám, hogy „te meg az a te hülye adatelemzésed” korlátozni fogja az unokái számát, a kutatásaim nem változtattak azon, hogy szeretnék gyerekeket. Ám még így is érdekes ez az nem helyénvaló gondolat, hogy valaki utólag megbánja a gyermekvállalást – olyan új aspektusát villantja fel az emberi természetnek, amely a hagyományos adathalmazokban nemigen érhető tetten. Kultúránk folyamatosan eláraszt minket a csodás, boldog családi élet képeivel. Az emberek zömének meg sem fordul a fejében, hogy a gyermekvállalást akár meg is lehet bánni. Pedig de. Ezt az ember talán senkinek se vallja be szívesen – kivéve a Google-t.

AZ IGAZSÁG A SZEXRŐL

Hány amerikai férfi meleg? A szexualitással kapcsolatos kutatások legendás kérdése ez. Mégis mind a mai napig az egyik legkeményebb dió a társadalomtudósok számára. A pszichológusok már nem hisznek Alfred Kinsey híres becslésének, mely szerint az amerikai férfiak 10 százaléka meleg, ugyanis a számítás alapjául szolgáló felmérésekben túltreprezentáltak a börtönviseltek és a prostituáltak. A mai kutatások 2-3 százalékra becsülik a melegek ará-

nyát. A szexuális preferencia azonban régóta olyan téma, amellyel kapcsolatban az emberek hajlamosak hazudni. Úgy gondolom, a Big Data segítségével a korábbiaknál pontosabb választ tudok adni erre a kérdésre.

Először is beszéljünk még egy kicsit a felmérések adatairól! A felmérések azt mutatják, hogy a toleráns államokban sokkal több a melegférfi, mint az intoleráns államokban. Egy Gallup-felmérés például kimutatta, hogy Rhode Island államban, ahol a legmagasabb a melegházasság támogatottsága, majdnem kétszer akkora a melegpopuláció számaránya, mint Mississippi államban, ahol a legalacsonyabb a melegházasság támogatottsága.

Ennek kétféle életszerű magyarázata lehet. Az első szerint az intoleráns államokban született melegférfiak a toleráns államokba költöznek. A második szerint az intoleráns államokban élő emberek nem feltétlenül fecsegik ki, hogy melegek, vagyis ezeken a helyeken még inkább hajlamosak hazudni, mint máshol.

Az első, a meleg mobilitásáról szóló magyarázat ügyében sok érdekes felismeréssel gazdagíthat minket egy másik Big Data-forrás: a Facebook, ahol a felhasználók megjelölhetik, hogy melyik nem iránt érdeklődnek. A nemi preferenciájukat megadó férfi facebookozók 2,5 százaléka állítja azt, hogy a férfiak érdeklik, ami hozzátétőlegesen megegyezik azzal, amit a felmérések eredményei jeleznek. És a Facebook is nagy különbséget mutat a magas, illetve alacsony toleranciaszintű államokban élő melegpopuláció lélekszámában: Rhode Islanden a Facebook szerint is kétszer akkora a melegférfiak aránya, mint Mississippiben.

A Facebook az emberek költözési szokásairól is szolgál információval. Kódolni tudtam például egy nyíltan melegfelhasználókból álló minta tagjainak lakóhelyét, így közvet-

lenül megbecsülhettem, hogy hány melegférfi költözik át az intoleráns államokból az ország toleránsabb vidékeire. A válasz? Egyértelmű, hogy van bizonyos fokú mobilitás – sokan költöznek például Oklahoma Cityből San Franciscóba. De számításaim szerint azok a férfiak, akik öszszepakolják a Judy Garland-CD-eket, hogy más, nyitottabb szellemiségű helyre tegyék át a székhelyüket, a toleráns, illetve intoleráns államok melegpopulációja közötti különbségnek csak kevesebb mint a felére adnak magyarázatot.*

Ráadásul a Facebook azt is lehetővé teszi, hogy közelebbről szemügyre vehessük a középiskolásokat. Rendhagyó csoport ez, amennyiben középiskolás fiúk elég ritkán választhatják meg szabadon a lakhelyüket. Ha tehát a nyíltan melegpopuláció számarányának államonkénti különbségeit a mobilitással magyarázzuk, akkor az lenne a logikus, ha a középiskolás felhasználók esetében nem mutatkozna ekkora különbség. Mit mondanak tehát a középiskolásokra vonatkozó adatok? Azt, hogy az intoleráns államokban sokkal kevesebb középiskolás fiú vállalja fel nyíltan, hogy meleg: Mississippiben például minden kétezredik. Tehát a különbség biztosan nem csak a mobilitásról szól.

Ám ha minden államban nagyjából ugyanannyi melegférfi születik, és a mobilitás sem nyújt kielégítő magyarázatot arra, hogy miért sokkal több egyes államokban a nyíltan melegférfi, akkor nyilván komoly szerepet játszik

* Egyesek sértőnek találhatják, hogy ha egy férfi Judy Garlandot hallgat, abból arra asszociálok – akár csak viccesen is –, hogy jobban szeret férfakkal szexelni. Természetesen nem azt akarom sugallni, hogy minden melegférfi, vagy akár a többségük, odavan a dívákért. Ugyanakkor a keresési adatokból kitűnik, hogy ebben a sztereotípiában van valami. Számításaim szerint egy olyan férfi, aki Judy Garlandról keres információt, mintegy háromszor akkora eséllyel néz melegpornót, mint heteró pornót. A Big Data arról árulkodik, hogy bizonyos sztereotípiák tényleg igazak.

a bujkálók létszáma. Már is visszakanyarodhatunk tehát a Google-hoz, amellyel annyi mindent megosztanak az emberek.

Van esetleg mód arra, hogy a pornókeresések alapján megvizsgáljuk, hány melegférfi él *valójában* az egyes államokban? Van bizony. A Google-keresések, illetve a Google AdWords adatai alapján úgy becsültem, hogy a férfiak által indított pornókereséseknek országos szinten mintegy 5 százaléka irányul férfi melegpornóra. (Ebbe ugyanúgy beleértendő a Rocket Tube nevű népszerű melegpornóoldalra és hasonlókra irányuló keresések is, mint a „melegpornó” keresőszó.)

És vajon hogyan változik ez az arány országgrészenként? Összességében elmondhatjuk, hogy a toleráns államokban több a melegpornós keresés, mint az intoleráns államokban. Ennek van értelme, hiszen a melegférfiak egy része, ugyebár, átköltözik a toleránsabb államokba. De korántsem akkora a különbség, amekkorának a felmérések vagy a Facebook szerint lennie kellene. Mississippi államban például számításaim szerint a pornóra rákereső férfiak 4,8 százaléka keres melegpornót, ami sokkal több, mint amennyi a felmérések vagy a Facebook adataiból következne, és ez az arány elég közel áll a Rhode Island-i adathoz (5,2%).

Hány amerikai férfi meleg tehát? A férfiak által indított pornókeresések között a mérés szerint durván 5 százalék a melegpornó aránya, ami reális becslésnek tűnik az Egyesült Államokban élő meleg lélekszámára vonatkozóan. És van egy másik, kevésbé nyilvánvaló módja is annak, hogy eljussunk ugyanehhez a számhoz. Csak egy kis adat-tudományra van hozzá szükségünk. A tolerancia szintje, illetve a nyíltan melegpopuláció számaránya közti összefüggést aknázhathatjuk ki. Csak egy kicsit legyenek még türelemmel.

Előzetes kutatásaim azt jelezték, hogy egy adott államban a melegházasság támogatottsága 20 százalékpontonként mintegy másfélszeresére növeli azon férfiak számát, akik nyíltan felvállalják a Facebookon, hogy melegek. Ebből kiindulva azt is megbecsülhetjük, hogy egy elképzelt, abszolút toleráns államban, ahol az emberek 100 százaléka támogatja a melegházasságot, hányan lennének nyíltan melegek az ott született férfiak közül. Számításaim szerint mintegy 5 százaléka, ami szépen passzol a pornókeresésekkel kapcsolatos adatokhoz. Ahhoz, hogy abszolút toleráns környezetben nőjenek fel, talán a kaliforniai Öbölvidéken élő középiskolás fiúk állnak a legközelebb. Közülük mintegy 4 százalék vállalja fel nyíltan a Facebookon, hogy meleg. Úgy tűnik, ez az adat is összecseng a számításaimmal.

Meg kell itt említenem, hogy a nőknél még nem sikerült megbecsülnöm az azonos neműekhez való vonzódás számarányát. A pornóadatoknak itt kevesebb hasznát veszi az ember, mivel sokkal kevesebb nő néz pornót, mint férfi, így a minta kevésbé reprezentatív. Ráadásul úgy tűnik, hogy a pornófogyasztó nők közül még azok is kedvelik a lesbikus pornót, akik a való életben a férfiakhoz vagy főként a férfiakhoz vonzódnak. A PornHubon nők által megtekintett videók nem kevesebb mint 20 százaléka lesbikus pornó.

Az, hogy az amerikai férfiak öt százaléka meleg, természetesen csak becslés. Vannak biszexuális férfiak is, ahogy olyanok is – különösen a fiatalok között –, akik nem tudják biztosan, hogy kihez is vonzódnak. Őket nyilván nem lehet olyan precízen megszámolni, mint a szavazókat vagy a mozilátogatókat.

Egyvalami viszont biztosan következik a számításaimból: az, hogy még mindig rengeteg amerikai bujkál, különösen az intoleráns államokban. Ők nem árulják el szexuális preferenciájukat a Facebookon. Nem vallják be azt

a felméréseken. És sokszor még az is előfordul, hogy nőkkel élnek házasságban.

Az is kiderül az adatokból, hogy a nők elég sűrűn gyanakodnak arra, hogy a férjük meleg. Ezt tükrözi „a férjem meleg?” keresőkifejezés meglepő gyakorisága. A „meleg” szó 10 százalékkal gyakoribb automatikus folytatása az „a férjem...” kezdetű kérdéseknek, mint a második helyezett „megcsal”. Valamint nyolcszor gyakoribb, mint az „alkoholista”, és tízszer gyakoribb, mint a „depressziós”.

A legárpokodóbb adat talán az, hogy a férj szexualitását megkérőjelező keresések a legkevésbé toleráns régiókban a leggyakoribbak. Az erre rákereső nők aránya Dél-Karolinában és Louisianában a legmagasabb. Ami azt illeti, a 25 állam közül, ahol a legtöbbször felteszik ezt a kérdést, 21-ben az országos átlag alatti a melegházasság támogatottsága.

A Google és a pornóoldalak nem az egyedüli adatforrások a férfiak szexualitására vonatkozóan. Máshonnan is megtudhatjuk, mit is jelent az, amikor valaki bujkál. Elemeztem a Craigslistre feltett apróhirdetéseket, amelyekben férfiak keresik az „alkalmi találka” lehetőségét. Megállapítottam, hogy a kevésbé toleráns államokban jellemzően magasabb az olyan férfiak százalékos aránya, akik férfikkal szeretnének alkalmi kapcsolatot létesíteni. Azon államok között pedig, ahol az ilyesmi százalékosan a leggyakoribb, ott találjuk Kentuckyt, Louisianát és Alabamát.

Most pedig, hogy még jobban bepillantassunk a bujkálók világába, térjünk vissza a Google-keresések adataihoz, és vesszünk el egy kicsit a részletekben. A „melegpornó” előtt vagy után beírt leggyakoribb keresőkifejezések egyike a „melegteszt”. (Az efféle tesztek készítői nagy mértéken arra vállalkoznak, hogy eldöntsék, egy férfi homoszexuális-e, vagy sem.) A „melegteszt” keresőszó is durván

kétszer olyan gyakori a legkevésbé toleráns államokban, mint másutt.

Vajon mit jelent az, hogy valakinél egymást követik a „melegpornó” és a „melegteszt” keresőszavak? Feltehetően azt, hogy az illető eléggé össze van zavarodva, vagy egyenesen háborog a lelke. Észszerű a feltevés, hogy egy részük megerősítésre vágyik: hátha a melegpornó iránti érdeklődése még nem azt jelenti, hogy ő is meleg.

A Google-keresésekre vonatkozó adatok nem engednek bepillantást egy-egy konkrét felhasználó meghatározott időtávra vonatkozó böngészési előzményeibe. Az AOL azonban 2006-ban kiadott a kutatóknak egy, a felhasználók kereséseit tartalmazó mintát. Íme, az egyik anonim felhasználó néhány keresése, amelyeket egy hatnapos időszakban indított:

Péntek 03:49:55 ingyenes meleg képek [sic]
Péntek 03:59:37 öltöző meleg képek
Péntek 04:00:14 meleg képek
Péntek 04:00:35 melegszex képek
Péntek 05:08:23 hosszú meleg kvíz
Péntek 05:10:00 jó melegteszt
Péntek 05:25:07 melegtesztetek összezavarodott férfinak
Péntek 05:26:38 melegtesztetek
Péntek 05:27:22 meleg vagyok-e tesztek
Péntek 05:29:18 meleg képek
Péntek 05:30:01 meztelen férfi képek
Péntek 05:32:27 ingyenes meztelen meleg képek
Péntek 05:38:19 szexi meleg szex képek
Péntek 05:41:34 szexi férfi seggbeszex
Szerda 13:37:37 meleg vagyok-e tesztek
Szerda 13:41:20 melegtesztetek
Szerda 13:47:49 szexi férfi seggbeszex
Szerda 13:50:31 ingyenes meleg szexvidió [sic]

Ez mindenképpen olyan férfit sejtet, aki nincs kibékülve a szexualitásával. A Google-adatok pedig arról árulkodnak, hogy még mindig sokan járnak hasonló cipőben, és a nagy többségük olyan államokban él, ahol kevésbé tolerálják az azonos neműek párkapcsolatát.

Hogy még közelebbi információkat szerezhessek a számok mögött álló emberekről, egy Mississippi államban élő pszichiáterhez fordultam, aki arra specializálódott, hogy bujkáló melegférfiaknak nyújt segítséget. Megkértem a szóban forgó pszichiátert, hogy kérdezze meg a pácienseit, nem lenne-e kedve valamelyiküknek beszélni velem. Egy férfi jelentkezett. Elmondta, hogy hatvanas éveiben jár, nyugalmazott egyetemi professzor, és több mint negyven éve él házasságban ugyanazzal a nővel.

Úgy tíz éve, amikor már nem bírta tovább az elfojtott szexualitásával járó stresszt, elment a pszichiáterhez, és végre felvállalta a szexuális orientációját. Addig is tudta, hogy vonzódik a férfiakhoz, de azt hitte, hogy ez általános dolog, amit a férfiak egyszerűen eltitkolnak. A terápia megkezdése után került sor élete első és egyetlen melegakutására: egy húszas éveiben járó diákjával volt együtt, és elmondása szerint „csodálatos” élményben volt része.

A feleségével nem élnek nemi életet. A professzor elmondta: büntudata lenne, ha valaha is véget vetne a házasságuknak, vagy nyíltan viszonyt kezdene egy férfival. Élete szinte minden fontos döntését megbánta.

A nyugalmazott professzor és neje éjszakai romantika nélkül, szerelem nélkül, szex nélkül telnek. Az utóbbi időben rengeteget javult a helyzet, de a szívós, makacs intolerancea jóvoltából még mindig több millió amerikai kénytelen így élni.

Olvasóimat talán nem sokkolja, hogy a férfiak 5 százaléka meleg, illetve hogy sokan mindmáig nem vállalják fel,

hogy azok. Ám volt idő, amikor az emberek zömét sokkolta volna a hír. És vannak olyan helyek, ahol a nagy többség ma is megdöbbenne.

„Irán nem olyan, mint a maguk országa, nálunk nincsenek homoszexuálisok” – jelentette ki határozottan Mahmúd Ahmadinezsád, Irán akkori elnöke 2007-ben. „Iránban ez a jelenség nem létezik.” Anatolij Pahomov, az oroszországi Szocsi polgármestere pedig nem sokkal az előtt, hogy a város otthont adott a 2014-es téli olimpiának, ezt mondta a melegekről: „A mi városunkban nincsenek ilyenek.” Az internet azonban arról tanúskodik, hogy mind Szocsiban, mind Iránban jelentős az érdeklődés a melegpornó iránt.

Ami felvet egy kézenfekvő kérdést: létezik olyan gyakori szexuális hajlam, amely az Egyesült Államokban is mindmáig sokkolónak számít? Hát, attól függ, mit minősítünk gyakorinak, és milyen könnyen sokkolódunk.

A Pornhubon leggyakrabban beírt keresőkifejezések nagy része nem hat a meglepetés erejével – a férfiak esetében olyanokat találunk közöttük, mint „tini”, „hármásban” és „szopás”, a nőkben pedig olyanokat, mint „szexuális szerelmeskedés”, „mellbimbószopogatás” és „férfi puncit nyál”.

Ha kívül merészkedünk a mainstreamen, a Pornhub-adatok olyan fétisekről is tudósítanak, amelyekről talán azt sem sejtettük, hogy léteznek. Akadnak nők, akik olyanokra keresnek rá, mint „alma análishan” vagy „plüssállatokkal dugni”. Akadnak férfiak, akik olyanokra keresnek rá, mint „takonyfétis” vagy „meztelen keresztrefeszítés”. De az efféle keresések ritkák: havonta talán ha tíz ilyen előfordul a Pornhub hatalmas online birodalmában.

A Pornhub-adatok áttekintése egy további, idekapcsolódó tanulságra is fényt derít: azt, hogy mindenkire vár va-

laki. A nők, nem túl meglepő módon, sokszor keresnek rá „magas”, „kreol bőrű” és „jóképű” fiúkra. De olykor „alacsony”, „sápadt” és „csúnya” srácokra is rákeresnek. Olyan nők is akadnak, akik „mozgássérült” fiúra, „kis farkú puffók sráca” vagy „ronda kövér öregemberre” keresnek rá. A férfiak gyakran keresnek „vékony” nőkre „nagy mellel”, valamint „szőke hajú” nőkre. De olykor „kövér” nőkre, „apró mellű” nőkre és „zöld hajú” nőkre is rákeresnek. És olyan férfiak is akadnak, akik „kopasz” nőkre, „törpe” nőkre vagy „mellbimbó nélküli” nőkre keresnek rá. Ezek az adatok igencsak örömteliek lehetnek azok számára, akik nem magasak, kreol bőrűek és jóképűek, illetve nem vékonyak, nagymellűek és szőkék.*

De mi a helyzet más, szintén gyakori és ugyancsak meglepő keresésekkel? A férfiak kereséseinek 150-es toplistájából engem személy szerint a Freudról szóló fejezetben már tárgyalt, vérfertőző témájú videók leptek meg a legjobban. A férfivágatok titokzatos, keveset tárgyalt tárgya még a „shemale”, azaz „hímnő” (77. leggyakoribb keresés), valamint a „granny”, azaz „nagyi” (110. leggyakoribb keresés). Összességében a férfiak Pornhub-kereséseinek mintegy 1,4 százaléka irányul pénisszel rendelkező nőkre, és 0,6 százaléka (a 34 év alatti férfiak esetében 0,4 százaléka) az idősebb korosztályra. Kifejezetten a serdületlen gyerekekre csak minden huszonnégyezredik, férfiak által indított Pornhub-keresés irányul. Ennek persze lehet valami kö-

* Úgy vélem, ezek az adatok az optimális randistratégia kialakításában is segítségünkre lehetnek. A dolog nyilván úgy működik, hogy az ember kimerészkedik, kap egy csomó visszautasítást, és ezeket nem veszi túlzottan magára. Ez az eljárás önt is hozzásegítheti, hogy végül megtalálja a társát, aki épp az olyanokhoz vonzódik a legjobban, amilyen ön. Még egyszer mondom: bárhogyan nézzen is ki, léteznek ilyen emberek. Nekem elhiheti.

ze ahhoz, hogy a Pornhub, nyilvánvaló okokból, tiltja a gyermekpornográfia minden formáját, ilyesmi birtoklása pedig törvényellenes.

A nők leggyakoribb Pornhub-keresései között akad a pornográfiának egy olyan műfaja is, amely, figyelmeztetem önöket, alkalmas lehet sokak nyugalmanak megzavarására: a nők elleni erőszakkal párosuló szex. A heteró pornóra való női keresések nem kevesebb mint 25 százaléka hangsúlyozza a nő fájdalmát és/vagy megaláztatását – úgymint „fájdalmas anális sírás”, „nyilvános megszégyenülés” vagy „extrém brutális csoportos szex”. Öt százalék a nem konszenzuális szexet – „nemi erőszakot” vagy „kényszert” – keres, még úgy is, hogy az ilyen videók le vannak tiltva a Pornhubon. Az ilyen tartalmú kifejezésekre való keresés legalább kétszer olyan gyakori a nőknél, mint a férfiaknál. Az olyan pornóműfajok tehát, amelyek nő ellen elkövetett erőszakot mutatnak be, az adatelemzésem alapján majdnem mindig aránytalanul nagyobb arányban vonzzák a nőket.

Természetesen fontos, hogy miközben próbáljuk feldolgozni az imént elmondottakat, arról se feledkezzünk meg, hogy a fantázia és a való élet két különböző dolog. Igen, a női nem azon kisebbségén belül, amely látogatja a Pornhubot, valóban van egy alcsoport, amely nemi erőszakot megjelenítő képsorokat keres – egyébként hiába. Hadd legyek egy kicsit szájbarágós: ez nem azt jelenti, hogy a nők a való életben is szeretnék, ha megerőszakolnák őket, ahogy az is nyilvánvaló, hogy ettől a tényről a nemi erőszak nem lesz kevésbé förtelmes bűn. Azt viszont bizvást leszűrhetjük a pornóadatokból, hogy az embereknek néha vannak olyan ábrándjaik, amelyeknek ők maguk sem örülnek, és amelyekről jó eséllyel soha nem tennének említést mások előtt.

A „szekrény”, amelyben az ember bujkál, nem csupán az ábrándok tárháza. Az emberek sok más titkot is rejtegetnek a szex tárgykörében – például azzal kapcsolatban, hogy mennyi jut nekik belőle.

A bevezetőben már esett róla szó, hogy az amerikaiak sokkal több óvszer elhasználásáról számolnak be, mint amennyi az eladási mutatókban szerepel. Persze ebből akár azt is leszűrhetnénk, hogy valójában nem szexelnek annyiszor *óvszerrel*, ahányszor mondják. Az adatokból azonban kitűnik, hogy már a szex gyakoriságát is eltülozzák. Például a 15 és 34 év közötti nők mintegy 11 százaléka állítja magáról, hogy szexuálisan aktív, jelenleg nem terhes, és nem használ fogamzásgátló eszközt. Ennek alapján a tudósok még akkor is azt várnák, hogy ennek a 11 százaléknak havonta a 10 százaléka teherbe essen, ha viszonylag konzervatívan becsüljük meg, hogy ez a szexuális aktivitás pontosan hány alkalmat is jelent. Csakhogy ez a 10 százalék máris több lenne, mint az összes terhesség száma az Egyesült Államokban (113 szülőképes korú nő közül 1 esik teherbe). Csak hát a mi szexmániás kultúránkban olykor nem könnyű beismerni, hogy az ember nem csinálja túl sokszor.

Ha viszont valaki megértő szavakra vagy éppen tanácsra vágyik, akkor működésbe lép a már említett ösztönző – vagyis elmondja a Google-nak, ami a szívét nyomja. A keresőmotornak tizenhatszor annyian panaszozzák el, hogy a házastársuk nem akar szexelni, mint azt, hogy nem akar velük beszélgetni. A nem házasságban élő partnerek közül öt és félszer annyian panaszkodnak arra, hogy a másik nem akar szexelni, mint arra, hogy nem válaszol az sms-eikre.

A Google-keresések pedig meglepő bűnös felett mondják ki az ítéletet a rengeteg szexmentes párkapcsolat ügyé-

ben. Kétszer annyi panasz érkezik arra, hogy a barát nem akar szexelni, mint arra, hogy a barátnő nem akar szexelni. A nők panaszai között utcahosszal vezet az, hogy „a barátom nem szexel velem”. (A Google-keresések nincsenek lebontva nemek szerint, de mivel az előző alfejezetben kiderült, hogy a férfiak 95 százaléka heteró, „a barátom...” kezdetű kereséseket aligha írják be túl sokszor férfiak.)

Hogyan értelmezzük ezt? Tényleg azt jelentené ez, hogy a férfiak többször tagadják meg a partnerüktől a szexet, mint a nők? Nem feltétlenül. Amint azt korábban már említettem, a Google-keresések torzíthatnak az olyan témák javára, amelyekről nemigen beszélnek az emberek, mert feszélyezve érzik magukat tőlük. A férfiaknak talán kevésbé kínos, amikor a barátnőjük szexuális érdeklődésének hiányáról mesélnek a barátaiknak, mint a nőknek, hogy kitérgetik a fiújuk passzivitását a barátnőikkel. Ám ha a Google-adatok nem arról tanúskodnak is, hogy a férfiak kétszer nagyobb valószínűséggel kerülnek a szexet, mint a barátnőik, azt mindenképpen sejtetik, hogy gyakrabban kerülnek, mint azt bevallanák.

A Google-adatok továbbá a szex gyakori kerülésének egyik okát is sejtetik: az irtózatosszorongást, amelynek jelentős része ráadásul teljességgel alaptalan. Kezdjük a férfakkal! Ők, ugyebár, sokat görcsölnek azon, hogy kellően jól vannak-e „felszerszámozva”. Ez nem újdonság, a szorongás mértéke viszont talán igen, ugyanis ez az aggály mélyreható és mindent átható.

A férfiak többször keresnek rá a nemi szervükre a Google-on, mint bármely más testrészükre: többször, mint a tüdejükre, a májukra, a lábukra, a fülükre, az orrukra, a torkukra és az agyukra együttrve. A férfiak többször keresnek rá arra, hogy mitől lehetne nagyobb a péniszük,

mint arra, hogy hogyan kell gitárt hangolni, omlettet sütni, kereket cserélni. A férfiak leggyakoribb, keresésben kifejezett aggálya a szteroidokkal kapcsolatban nem az, hogy nem károsak-e általában véve az egészségre, hanem az, hogy nem zsugorítják-e össze a hímtagjukat. A férfiak leggyakrabban keresett kérdése a testükben, illetve az elméjükben a kor előrehaladtával bekövetkező változások kapcsán ugyancsak az, hogy nem lesz-e kisebb a péniszük.

Széljegyzet: a férfiak nemi szervével kapcsolatos egyik leggyakoribb Google-keresés az, hogy „mekkora a péniszem?”. Azt hiszem, az, hogy a férfiak ezzel kapcsolatban inkább fordulnak a Google-hoz, mint egy vonalzóhoz, jól fejezi ki a mi digitális korunk lényegét.*

Érdekl-e a nőket a pénisméret? A Google-keresések tanúsága szerint csak igen ritkán. Minden olyan keresésre, amelyet egy nő ír be a partnere falloszára vonatkozóan, mintegy 170 olyan keresés jut, amelyet egy férfi indít a sajátjáról. Igaz, azon ritka esetekben, amikor a nők aggodalmukat fejezik ki a partnerük péniszé kapcsán, elég sűrűn esik szó annak méretéről, de nem feltétlenül azért, mert kicsi. A panaszok több mint 40 százaléka arra vonatkozik, hogy a szóban forgó pénisz túl nagy. A „...szex közben” szavakkal végződő keresések elején a legtöbbször a „fájdalom” szó szerepel (az első ötbe „vérzés”, a „piselés”, a „sírás” és a „fingás” került még be). A férfiak mérettel kapcsolatos kereséseinek mégis mindössze 1 százaléka szól arról, hogy mitől lehetne kisebb a péniszük.

* Eredetileg azt a címet akartam adni a könyvnek, hogy: *Mekkora a péniszem? Mit tudhatunk meg a Google-keresésekből az emberi természetről?* De a szerkesztőm figyelmeztetett, hogy egy ilyen könyvet kemény kihívás lenne eladni. Szerinte az embereknek kínos lenne megvenni egy könyvet ezzel a címmel, mondjuk, a reptéri könyvesboltban. Önök szerint is?

A férfiak második leggyakoribb kérdése a szexszel kapcsolatban az, hogy miként hosszabbíthatnák meg az együttléteket. Úgy tűnik, hogy a bizonytalanságaik itt sem vágnak egybe a nők aggályaival. A nők körülbelül ugyanannyiszor kérdeznek rá arra, hogy miként juttathatnák előbb a csúcsra a barátjukat, mint arra, hogy miként érhetnék ezt el lassabban. A barátjuk orgazmusával kapcsolatos leggyakoribb aggály pedig nem is az, hogy mikor sikerül elérni, hanem az, hogy miért nincs.

A férfiak kapcsán nem sűrűn esik szó a testképpel kapcsolatos problémákról. Pedig, noha igaz, hogy a külső iránti érdeklődés nagyobb mértékben jellemző a nőkre, annyira azért nem billen felénk a mérleg nyelve, amennyire a sztereotípiák sugallják. A Google AdWords adataival elvégzett elemzésem szerint – ez az alkalmazás méri, hogy milyen weboldalak látogatnak az emberek – a szépség és a fitness iránt 42 százalékban, a fogyás iránt 33 százalékban, a kozmetikai sebészet iránt pedig 39 százalékban férfiak érdeklődnek. A „hogyan kell”-lel kezdődő, melleskel kapcsolatos kereséseknek is mintegy 20 százalékát teszi ki az a kérdés, hogy hogyan lehet megszabadulni a lötytyedt férfimelltől.

De ha az e téren nem túl magabiztos férfiak száma magasabb is, mint gondolnánk, a nők így is simán lekörözik őket a kinézetük miatti görcsölésben. Mit árul el tehát a digitális igazságszérum a nők önmagukkal kapcsolatos aggályairól? Például azt, hogy az Egyesült Államokban évente több mint hétmillió keresés irányul a mellimplantátumokra, pedig a hivatalos statisztika szerint csak mintegy 300 ezer nő csinálja végig a procedúrát.

A nőknek a saját hátsójukhoz fűződő viszonyát ugyancsak nagyfokú bizonytalanság jellemzi. Igaz, nemrég so-

kuk agyában éppen az ellenkezőjére fordult az, hogy miért nem tetszik nekik a fenékük.

2004-ben a fenékmódosításra irányuló keresések többször a kisebbítésre irányultak. Az a kívánság, hogy valakinek nagyobb legyen a fenéke, főként a feketék lakta térségekre koncentrálódott. 2010 óta azonban az Egyesült Államok többi részén is gyakoribbá vált a nagyobb fenék iránti igény. Ha maga a fenékméret nem is, az ilyen irányú érdeklődés mértéke négy év alatt megháromszorozódott. 2014-ben már minden államban többször kerestek rá a nagyobbításra, mint a kisebbítésre. Manapság pedig minden ötödik mellimplantátum-keresésre jut egy fenékimplantátum-keresés. (Köszönjük, Kim Kardashian!)

De vajon egybevág-e a nők egyre intenzívebben terjedő nagyfenék-preferenciája a férfiak preferenciáival? Érdekes módon igen. A „nagy segg pornó” keresőkifejezés, amely azelőtt szintén a fekete közösségekben koncentrálódott, mára az egész országban sokkal népszerűbbé vált.

Mire vágnak még a férfiak a női test kapcsán? Amint azt már említettem – és amint az a többség számára ordítóan nyilvánvaló –, a nagy melleket részesítik előnyben. A meghatározott tartalmú pornókeresések 12 százaléka nagymellű nőkre irányul. Ez csaknem hússzorosa a kis-mellűekre irányuló keresések mennyiségének.

Mindezek ellenére nem egyértelmű, hogy a férfiak olyan nőre vágnának, akinek mellimplantátuma van. A nagymellű nőkre irányuló pornókeresések 3 százalékában kifejezetten kikötik, hogy természetes melleket akarnak látni.

A felhasználók feleségéről és a mellimplantátumról szóló Google-keresések nagyjából fele-fele arányban kíváncsiak arra, hogy mivel vehetnék rá az asszonyt a műtetre, illetve arra, hogy ugyan minek az neki.

Vagy vegyük a barátnő mellével kapcsolatos leggyako-

ribb keresést: „imádom a barátnőm mellét”. Nem egészen világos, mit is remélnek a férfiak a Google-tól egy ilyen keresés nyomán.

Ahogy a férfiaknak, úgy a nőknek is vannak kérdéseik a nemi szervükkel kapcsolatban. Sőt mi több, közel ugyanannyi kérdést tesznek fel a hüvelyükről, mint a férfiak a péniszükről. A hüvelyükkel kapcsolatos aggályaik sokszor egészségügyi jellegűek, de a kérdések legalább 30 százaléka egészen másról szól. A nők tudni akarják, hogyan borotválják a nemi szervüket, mitől lesz feszesebb vagy éppen jobb ízű. Valamint – ahogy arról már szó esett – az is döbbenetesen gyakori aggály, hogy mitől lehetne illatosabb.

A nőket a legtöbbször az aggasztja, hogy a hüvelyük halszagú. Ezt az ecetszag, a vöröshagymaszag, az ammóniaszag, a fokhagymaszag, a sajtíz, a testszag, a pisiszag, a kenyérszag, a klórszag, az ürülékszag, az izzadságszag, a fémszag, a lábszag, a szemétszag és a rothadó hús szaga követi a sorban.

A férfiak általában nem sűrűn keresnek rá a Google-on a partnerük nemi szervére. Durván ugyanannyiszor keresnek rá a barátnőjük hüvelyére, mint a nők a barátjuk péniszére.

Amikor viszont rákeresnek a partnerük hüvelyére, rendszerint arra panaszkodnak, ami a nőket a legjobban aggasztja: a szagára. A férfiak többnyire arra kíváncsiak, hogy miként hozhatnák szóba a témát úgy, hogy ne bántásuk legyen a nőt. De az is előfordul, hogy a hüvely szagára vonatkozó kérdéseik a saját bizonytalanságukat tükrözik. Néha arra kérdeznak rá, leleplezhetik-e a hüvely szaga alapján, ha csalják őket? Nyomra vezetheti-e őket az óvszárszag, illetve egy másik férfi ondójának a szaga?

Mihez kezdjünk ezzel a sok titkos bizonytalansággal? Itt egyértelműen jó hírrel szolgálhatok. A Google számos

nyomás okot ad arra, hogy kevesebbet aggódjunk, mint amennyit szoktunk. Az azzal kapcsolatos, mélyen gyökeröző félelmeink, hogy milyenek érzékel minket a partnerünk, sokszor alaptalanok. A partnerekről, ha egyedül ülnek a számítógép előtt, és semmi sem ösztönzi őket arra, hogy hazudjanak, kiderül, hogy elnézőek, és távolról sem annyira felszíneseek. Ami azt illeti, annyi időt töltünk az-
zal, hogy a saját testünk felett ítélgessünk, hogy a mások teste feletti ítélgességre már nemigen marad energiánk...

Az is elég valószínű, hogy a szexuális témájú Google-keresésekben megjelenő két súlyos aggály: a szex hiánya, illetve a saját szexuális vonzerőnkkel és teljesítményünkkel kapcsolatos bizonytalanság nem független egymástól. A kettő talán összefügg. Ha kevesebbet görcsölnénk a szex miatt, még az is lehet, hogy többször csinálnánk.

Mit mondhatnak még nekünk a Google-keresések a szexről? Eljátszhatjuk például a nemek csatáját: megnézhetjük, ki a bőkezűbb, az adakozóbb. Vegyük például az olyan keresések számát, amelyek arra vonatkoznak, hogyan lehetnénk jobbak a másik nem orális kényeztetésében! A férfiak vagy a nők keresnek gyakrabban ilyen tippeket? A férfiak vagy a nők önzetlenebbek szexuálisan? Persze hogy a nők. Mindent egybevetve 2:1-re becsülöm az arányt a nők javára: kétszer annyiszor kérnek tanácsot a Google-tól arra vonatkozóan, hogy miként tudnák jobb orális szexben részesíteni a partnerüket.

Ráadásul a férfiak esetében, amikor tippeket keresnek arra, hogy miként kényeztessenek valakit orálisan, ez a valaki sokszor még csak nem is egy másik személy. A férfiak ugyanannyiszor keresnek rá önmaguk orális kielégítésére, mint arra, hogy miként juttassanak el az orgazmusig egy nőt. (Ez az egyik kedvenc Google-adatom.)

AZ IGAZSÁG A GYŰLÖLETRŐL ÉS AZ ELŐÍTÉLETEKRŐL

A szex és a szerelem aligha az egyedüli témák, amelyeket szégyenérzet és titkolózás övez. Sokan például – jó okkal – hajlamosak megtartani maguknak az előítéleteiket. Gondolom, már az is előrelépésnek nevezhető, hogy az amerikaiak jó része ma már úgy érzi: ha nem akarja, hogy elítéljék, jobb, ha nem ismeri be, hogy etnikai, vallási hovatartozás vagy szexuális orientáció alapján ítélgessék mások felett. Sokan ugyanakkor ezzel együtt is mindmáig így tesznek. (Ismét figyelmeztetnem kell olvasóimat, hogy a nyugalom megzavarására alkalmas tartalom következik!)

Tanú erre a Google, ahol a felhasználók időnként olyan kérdéseket tesznek fel, mint: „miért bunkók a feketék?“, illetve: „miért szemetek a zsidók?“. Az alábbi táblázatban felsorolom azt az öt negatív töltetű szót, amelyek a leggyakrabban fordulnak elő az egyes embercsoportokra irányuló keresésekben.

	1.	2.	3.	4.	5.
AFRO-AMERIKAIK	bunkó	rasszista	buta	ronda	lusta
ZSIDÓK	gonosz	rasszista	ronda	smucig	kapzsi
MUSZLIMOK	gonosz	terrorista	rossz	erőszakos	veszélyes
MEXIKÓIAK	rasszista	buta	ronda	lusta	ostoba
ÁZSIAIK	ronda	rasszista	irritáló	buta	smucig
MELEGEK	gonosz	megtévedt	buta	irritáló	önző
KERESZTÉNYEK	buta	őrült	ostoba	téveszmés	megtévedt

A sztereotípiák között tetten érhetők bizonyos szabályszerűségek. A „bunkó” jelzővel például egyedül az afroamerikaiaknak kell szembesülniük. A „buta” besorolást mindegyik csoport megkapja, a zsidókat és a muszlimokat kivéve. A „gonosz” jelzőt a zsidókra, a muszlimokra és a melegekre is rásütik, a feketékre, a mexikóiakra, az ázsiaiakra és a keresztényekre viszont nem.

Terroristaként egyedül a muszlimokat sztereotipizálják – és ha egy amerikai muszlim beigazolja ezt a sztereotípiát, a reakció villámgyors és heves. Hála a Google keresési adatainak, percről percre nyomon követhetjük a gyűlölettől fűtött indulatok tombolását.

Lássuk például, mi történt 2015. december 2-án, a kaliforniai San Bernardinóban lezajlott mészárlás után. A két, félautomata pisztolyokkal és gépkarabélyokkal felfegyverzett támadó, Rizwan Farook és Tashfeen Malik berontott Farook munkatársainak összejövetelére, és lemészárolt tizennégy embert. Este pedig, szó szerint percekkel az után, hogy a média először közölte a tettesek muszlimos hangzású nevét, Kaliforniában máris felkavaróan sok ember eldöntötte, mi a teendő a muszlimokkal: meg kell ölni őket.

Kaliforniában akkor a „muszlimok” szót tartalmazó leggyakoribb Google-keresés az volt, hogy „muszlimokat ölni”. Országos szinten nagyjából ugyanannyiszor kerestek rá az emberek a „muszlimokat ölni” kifejezésre, mint arra, hogy „martini recept”, „migrén tünetei” vagy „Dallas Cowboys összeállítás”. A San Bernardinó-i támadást követő napokban minden „iszlamofóbia” miatt aggódó amerikai rá jutott egy másik, aki arra keresett rá, hogy „muszlimokat ölni”. A támadás előtt a muszlimokkal kapcsolatos keresések mintegy 20 százalékát motiválta a gyűlölet, a támadás utáni órákban viszont a keresések mennyiségének több mint felét.

És ezek a percről percre nyomon követett keresési adatok arról is árulkodnak, hogy milyen nehéz ilyenkor lecsillapítani a kedélyeket. Obama elnök négy nappal a lövöldözés után, főműsoridőben szólt a nemzethez. Egyszerre igyekezett megnyugtatóan az amerikaiakat afelől, hogy a kormánygátat tud szabni a terrorizmusnak, és ami talán még fontosabb, ezt a veszélyes iszlamofóbiát is le tudja csendesíteni.

Obama a jelek szerint a jobbik énünkre próbált építeni: a befogadás és a tolerancia fontosságáról beszélt. Hatásos és megindító volt a retorikája. A *Los Angeles Times* meg is dicsérte, amiért arra figyelmeztetett: „ne hagyjuk, hogy a félelem elhomályosítsa az ítélőképességünket”. A *The New York Times* „keménynek”, ugyanakkor „megnyugtatónak” minősítette a beszédet. A Think Progress nevű hírportál is úgy méltatta, mint „a jó kormányzás szükséges eszközét, amely a muszlim amerikaiak életének megmentését célozza”. Vagyis Obama beszédét jelentős sikerként könyvelték el. De vajon tényleg az volt?

A Google keresési adatai szerint nem. Megvizsgáltuk az adathalmazt Evan Soltasszal, aki akkoriban a Princetonon dolgozott. Az elnök ezt mondta a beszédében: „Minden amerikainak vallásától függetlenül közös felelőssége, hogy elutasítsa a diszkriminációt”. Csakhogy a muszlimokat „terroristának”, „rossznak”, „erőszakosnak” és „gonosznak” minősítő keresések száma a beszéd ideje alatt, illetve közvetlenül utána megduplázódott. Obama elnök ezt is mondta: „A mi felelősségünk, hogy elutasítsuk a felekezeti tesztet annak kapcsán, hogy kit engedünk be az országba”. Csakhogy az akkoriban épp biztonságos menedéket kereső, javarészt muszlim vallású szír menekültekkel kapcsolatos, negatív tartalmú keresések száma 60 százalékkal nőtt, miközben az arra vonatkozó kérdések száma, hogy miként segíthetne valaki a szír menekülteken, 35 száza-

lékkel csökkent. Obama arra kérte az amerikaiakat, hogy „ne feledjék: a szabadság erősebb, mint a félelem”. A beszéd ideje alatt mégis megduplázódott a „muszlimokat ölni” keresőkifejezés gyakorisága. Sőt mi több, a beszéd alatt és után nagyjából minden elképzelhető negatív keresőkifejezésnek jelentősen megugrott a gyakorisága, miközben nagyjából minden elképzelhető pozitív keresése csökkent.

Másképp fogalmazva, úgy tűnt, Obama pontosan azt mondja, amit kell. A hagyományos média lelkesen gratulált is neki gyógyító hatású szavaihoz. Az internetről származó, digitális igazságszérumként működő, újszerű adatok viszont azt sugallták, hogy a beszéd, legalábbis annak elsődleges célja szempontjából, épp visszafelé sült el. Az internet tanúsága szerint nemhogy lecsendesítette volna a feldühödött tömeget, hanem sokkal inkább olajat öntött a tűzre. Az, amiről azt hisszük, hogy működik, olykor pontosan az ellenkező hatást érheti el. Olykor nagy szükségünk van az internetes adatokra, nehogy az ösztönünkre hallgatva akkor veregessük vállon magunkat, amikor nem kellene.

De akkor mit kellett volna mondania Obamának, hogy lecsillapítsa a gyűlöletnek ezt a formáját, amely manapság ennyire erősen jelen van Amerikában? Erre később még visszatérünk. Most vessünk egy pillantást az amerikai előítéletesség egyik ősrégi vonulatára: a gyűlöletnek arra a válfajára, amely kitűnik a többi közül, amely a legdestruktívabb, és amely a könyv elején ismertetett kutatás témáját is szolgáltatta. Amikor elkezdtem dolgozni a Google keresési adataival, az bizonyult a gyűlölettel kapcsolatos legárgulakodóbb ténynek, hogy az interneten mennyire népszerű a „nigger” szó.

Egyes számú vagy többes számú formában országos szinten évente hétmillió keresésben bukkan fel a „nigger”

szó. (Ismétlem: a rapszámokban szinte mindig a „nigga” szóalak fordul elő, így az eredményt nem befolyásolják jelentősen a hiphop-szövegek.) A „nigger viccek” keresőkifejezés tizenhétszer gyakoribb, mint a „kike (bibsi) viccek”, a „gook (koreai) viccek”, a „spic (latino) viccek”, a „chink (kínai) viccek” és a „fag (buzi) viccek” együttvéve.

Hogy mikor fordul elő leggyakrabban a „nigger(ek)”, illetve a „nigger viccek” keresőkifejezés? Amikor afroamerikaiak szerepelnek a hírekben. Az ilyen keresések száma többek között akkoriban volt a legmagasabb, amikor a Katrina hurrikánt követően a tévében és az újságokban túlélésért küzdő, kétségbeesett New Orleans-i feketéket lehetett látni. Obama első megválasztásakor is tetőztek. És a „nigger viccekre” való keresések száma átlagosan 30 százalékkal megugrik minden január harmadik hétfőjén, a Martin Luther King-émléknapon.

Az efféle, faji alapú gyalázkodás ijesztő elterjedtsége kétegyetket ébreszt a tekintetben, hogy nem sántít-e a rasszizmussal kapcsolatos jelenlegi felfogásunk.

Minden valamirevaló, a rasszizmusra vonatkozó elméletnek megoldást kell találnia Amerika egyik nagy rejtélyére. Egyrészt az amerikai feketék túlnyomó többsége úgy véli, elszenvedti az előítéletességet, amire bőven van bizonyítékuk is: a rendőri ellenőrzések során, az állásinterjúkon és az esküdtszékek döntéseiben tetten érhető diszkrimináció. Másrészt azonban csak nagyon kevés fehér amerikai ismeri be, hogy a maga részéről rasszista lenne.

A politikatudomány berkeiben az az uralkodó álláspont, hogy az ellentmondást a széles körben elterjedt *implicit* előítéletek magyarázzák. Az elmélet szerint a fehér amerikaiakban, minden jó szándék mellett, munkál egy tudatalatti elfogultság, amely befolyásolja a feketékkel szembeni

bánásmódjukat. A kutatók igen leleményes módszert találtak ki ennek az elfogultságnak a tesztelésére: az implicit asszociációs tesztet.

A vizsgálat egyértelműen igazolta, hogy az emberek nagy többségének néhány ezredmásodperccel hosszabb idő kell ahhoz, hogy fekete arcokhoz pozitív töltetű szavakat (például: „jó”) társítson, mint ahhoz, hogy negatívakat (például: „szörnyű”). Fehér arcoknál viszont fordítva működik a dolog. A pluszidő, amelyet a folyamat igénybe vesz, az implicit előítélet jelenlétének bizonyítéka – amelynek az alany talán nincs is tudatában.

Mindazonáltal a diszkriminációnak, amelyet a feketék érzékelnek, a fehérek pedig tagadnak, van egy alternatív magyarázata is: a rejtett *explicit* rasszizmus. Igen, feltehetően létezik egy viszonylag széles körben elterjedt, tudatos rasszizmus, amelyről az emberek nagyon is tudnak, de nem vallják be – pláne nem a közvélemény-kutatóknak. Úgy tűnik, a keresési adatok is erről árulkodnak. Semmi implicit nincs abban, amikor valaki „nigger vicceket” keres a neten. Azt is nehéz elképzelni, hogy amikor a felhasználók ugyanannyiszor keresnek rá a „nigger” szóra, mint a „migrén”-re és a „közgazdász”-ra, az afroamerikaiak életére nem lesz komoly hatással az *explicit* rasszizmus. A Google-adatok megjelenése előtt nem tudtuk meggyőzően mérni ezt az izzó ellenségességet. Most viszont már tudjuk. Így hát azt is módunkban áll megfigyelni, mi mindenre ad magyarázatot.

Például arra – amint arról korábban már szó esett –, hogy mi nyomta le az ország számos területén az Obamára leadott szavazatok mennyiségét a 2008-as, majd a 2012-es elnökválasztáson. Az adatok ezen kívül korrelálnak annak a jelentésnek a tartalmával is, amelyet a fehérek és feketék

közötti bérkülönbség tárgyában nemrégiben publikált egy közgazdász kutatócsoport. Egyszerűen szólva: azokban a régiókban, ahol a kutatásaim szerint a legtöbb volt a rasszista indíttatású Google-keresés, alulfizetik a feketéket. És aztán ott van még Donald Trump jelöltségének a kérdése is. A bevezetőben már említettem, hogy amikor Nate Silver, a közvélemény-kutató guru megvizsgálta, mely földrajzi változó korrelál a legerősebben Trumpnak a 2016-os republikánus előválasztáson mért támogatottságával, az általam kifejlesztett rasszizmustérképen találta meg a választ. Ez a bizonyos változó a „nigger(ek)” keresőszó volt.

Tudósok nemrégiben összeállították a feketékkel szembeni implicit előítélet államonkénti adatait. Ez lehetővé tette számomra, hogy összevessem a Google-keresések által mért *explicit* rasszizmus, illetve az *implicit* elfogultság hatásait. Egyebek mellett azt vizsgáltam, hogy mennyire dolgozott Obama ellen egyik, illetve másik tényező a két elnökválasztáson. A regresszióanalízis kimutatta, hogy ha ki akarom deríteni, hol teljesített gyengébben Obama, akkor az egyes régiókban indított rasszista Google-keresések száma sok mindenre magyarázatot ad. Egy-egy régiónak az implicit asszociációs teszteken elért eredményei viszont nem sokat tettek hozzá az összképhez.

Most pedig hadd terjesszek elő egy provokatív és további kutatásokra ösztönző sejtést, amely csak arra vár, hogy a különböző szakterületek művelői teszteljék. Az afroamerikaiakkal szembeni diszkrimináció elsődleges magyarázata eszerint nem az, hogy a laborkísérletekben önként részt vevő emberek tudat alatt negatív szavakat társítanak a feketékhez; hanem az a tény, hogy még mindig fehér amerikaiak milliói foglalatосkodnak olyan dolgokkal, mint például a „nigger viccek” keresése a neten.

Úgy tűnik, a feketék által rendszeresen megtapasztalt diszkriminációt inkább az explicit, bár rejtegetett rosszindulat táplálja. Más csoportokra viszont nagyobb hatással lehet a tudat alatti előítéletesség. A Google-keresések adathalmazának segítségével például meggyőződhettem a népesség egy másik csoportját, a kislányokat sújtó implicit előítéletek jelenlétéről.

Ugyan ki lenne elfogult a kislányokkal szemben? – kérdezhetik most önök.

Hát a szüleik.

Aligha meglepő, hogy a kisgyerekes szülőket izgalomba hozza a gondolat, hogy csemetéjük akár tehetséges is lehet. „A kétéves gyerekem/fiam/lányom...” kezdetű Google-keresések leggyakoribb folytatása az, hogy „tehetséges?”. Csakhogy ez a kérdés nem azonos arányban vonatkozik a kisfiúkra és kislányokra. A szülők két és félszer gyakrabban kérdezik, hogy „a fiam tehetséges?”, mint azt, hogy „a lányom tehetséges?”. Hasonlóan elfogultak akkor is, amikor másképp fogalmazzák meg a szellemi képességekre vonatkozó kérdéseiket, amelyeket hangosan szégyellnének is kimondani, például úgy, hogy „a fiam zseni?”.

Elképzelhető, hogy azért viselkednek így a szülők, mert valóban létező különbségeket érzékelnek a kisfiúk és a kislányok között? Lehet, hogy a kisfiúk többször használnak „nagy” szavakat, vagy többször mutatják a tehetség más, objektíven megfigyelhető jeleit, mint a kislányok? Hát nem. Sokkal inkább az ellenkezője igaz. A mérések egyértelműen azt mutatják, hogy a lányoknak kiskorukban nagyobb a szókincsük, és összetettebb mondatokban beszélnek. Az amerikai iskolákban a lányok arányosan 9 százalékkal nagyobb valószínűséggel kerülnek be valamilyen tehetséggondozó programba, mint a fiúk. Ennek ellenére a

vacsoraasztalnál körbepillantó szülők a jelek szerint tehetségesebbnek látják a fiúkat, mint a lányokat.* Mi több, az általam vizsgált, szellemi képességekre vonatkozó keresőkifejezések mindegyikét inkább a fiaik, mintsem a lányaik kapcsán tették fel a szülők, köztük azokat is, amelyek a szellemi képességek hiányára utaltak. Arra is több keresést találunk, hogy „a fiam visszamaradott?”, illetve „a fiam buta?”, mint ugyanezekre a kérdésekre a lányok kapcsán. A negatív töltetű keresőszavak esetében azonban, amilyen a „visszamaradott” vagy a „buta”, kevésbé billen a fiúk felé a mérleg nyelve, mint a pozitívakéban, amilyen a „tehetséges” vagy a „zseni”.

Akkor viszont mi a szülők legfőbb aggálya a lányukkal kapcsolatban? Mindenekelőtt bármi, ami a kislány külsejével kapcsolatos. Vegyük például a gyerekek testsúlyára vonatkozó kereséseket. A szülők mintegy kétszer gyakrabban keresnek rá arra, hogy „a lányom túlsúlyos?”, mint arra, hogy „a fiam túlsúlyos?”. Arról is kétszer annyszor érdeklődnek, hogy mivel vehetnék rá a lányukat a fogyásra, mint arról, hogy mivel érhetnék el ugyanezt a fiuknál. Pedig a testsúllyal is ugyanaz a helyzet, mint a tehetséggel: a nemek szerinti elfogultságnak nincs túl sok valós alapja. A lányok mintegy 28 százaléka túlsúlyos, míg a fiúknak mintegy 35 százaléka. De hiába jeleznek Amerika mérlegei több túlsúlyos fiút, mint lányt, a szülők több túlsúlyos lányt látnak – vagy legalábbis jobban aggasztja őket lányaik túlsúlya.

A szülők azt is másfélszer annyszor kérdezik meg, hogy szép-e a lányuk, mint azt, hogy jóképű-e a fiuk. Azt pedig,

* Hogy további teszteknek vethessem alá a hipotézisemet, gyereknevelési weboldalakról próbálok adatokat gyűjteni. Így sokkal nagyobb lesz a mintám, mint a mostani, amelyben egyelőre csak azok a szülők szerepelnek, akik beírták a Google-ba az említett keresőkifejezések valamelyikét.

hogy csúnya-e a lányuk, háromszor nagyobb valószínűséggel írják be a keresőbe, mint azt, hogy csúnya-e a fiuk. (Hogy honnan tudhatná a Google, hogy melyik gyerek mennyire szép vagy csúnya, azt nehéz lenne megmondani.)

Úgy tűnik, a szülők általánosságban is inkább használnak pozitív töltetű szavakat, amikor a fiaikról faggatják a Google-t. Hajlamosabbak megkérdezni, hogy boldog-e a fiuk, és kevésbé hajlamosak megkérdezni, hogy depressziós-e.

Liberális olvasóim most azt gondolhatják, hogy az efféle elfogultság inkább az ország konzervatívabb tájain jellemző, erre azonban semmiféle bizonyítékot nem találtam. Sőt, egyáltalán nem találtam szignifikáns kapcsolatot az efféle elfogultság, illetve egy-egy állam politikai vagy kulturális összetétele között. Ahogy arra sincs bizonyíték, hogy az efféle elfogultság mértéke csökkent volna 2004 óta, amikortól a Google keresési adatai a rendelkezésünkre állnak. Úgy tűnik, hogy a lányokkal szembeni elfogultság jóval elterjedtebb, és jóval mélyebben gyökerezik, mint azt hinni szeretnénk.

Nem a szexizmus az egyetlen téma, amelynek kapcsán az előítéletekre vonatkozó sztereotípiáink csődöt mondatnak.

Vikingmaiden88 huszonhat éves. Szívesen olvas történelmi tárgyú könyveket, és verseket is ír. A mottója egy Shakespeare-idézet. Mindezt a profiljából és a posztjaiból szűrtem le, amelyek Amerika legnépszerűbb online gyűlöletoldalán, a Stormfront.orgon található. Azt is megtudtam, hogy Vikingmaiden88-nak tetszett a *The New York Times* weboldalának a tartalma, az egyik újságé, ahol én is dolgozom. Lelkes hangvételi posztot írt a *Times* egyik cikkéről.

A közelmúltban több tízezer ehhez hasonló Stormfront-profil elemeztem. A regisztrált tagok bevihetik a lakhe-

lyüket, a születési dátumukat, az érdeklődési körüket és más személyes információkat.

A Stormfrontot 1995-ben alapította Don Black volt Ku Klux Klan-vezér. Az oldal két legnépszerűbb csoportját úgy hívják, hogy „Nemzetiszocialista Unió”, illetve „Adolf Hitler rajongói és támogatói”. A Quantcast szerint 200 ezer és 400 ezer közé tehető azok száma, akik az utóbbi egy év minden hónapjában ellátogattak a Stormfrontra. A Southern Poverty Law Center nevű jogvédő szervezet egyik legfrissebb jelentése azt állítja, hogy az utóbbi öt évben közel száz gyilkosság köthető a Stormfront tagjaihoz.

A Stormfront tagjai, mint kiderült, nem azok, akiknek hittem őket.

Jellemzően fiatalok, legalábbis a megadott születési dátumuk alapján. A legtöbb tag tizenkilenc éves korában lép be. És négyszer annyi tizenkilenc éves csatlakozik a Stormfronthoz, mint ahány negyvenéves. Tény, hogy az internet, illetve a közösségi oldalakat javarészt fiatalok használják, de azért nem ennyire elsőpró arányban.

A profilban nincs olyan mező, ahol a felhasználó megadhatná a nemét. Én azonban átnéztem a tagok egy véletlenszerű mintájának összes posztját és teljes profilját, és kiderült, hogy a tagok többségének ki lehet következtetni a nemét. Ennek alapján arra jutottam, hogy a Stormfront-tagok mintegy 30 százaléka nő.

Az egy főre jutó Stormfront-tagok száma Montanában, Alaszkában és Idahóban a legmagasabb. Ezen államok lakossága túlnyomórészt fehér. Vajon ez azt jelenti, hogy ha valaki kevésbé sokszínű környezetben nő fel, az termékeny táptalaja a gyűlöletnek?

Valószínűleg nem. Inkább az lehet a háttérben, hogy mivel a felsorolt államokban magasabb a nem zsidó fehérek aránya, a zsidókat és nem fehéreket támadó csoportok

potenciális tagjainak létszáma is eleve nagyobb. Valójában a Stormfront célközönségéhez tartozók nagyobb százalékból lépnek be a szervezetbe olyan térségekben, ahol több kisebbségi él. Különösen igaz ez a tizennyolc éven aluli Stormfront-tagokra, akik nem választhatják meg önállóan a lakóhelyüket.

Kaliforniában, ahol a legmagasabbak közé tartozik a kisebbségi populáció számaránya, ez a korosztály az országos átlagnál 25 százalékkal nagyobb arányban képviselteti magát a tagság körében.

A weboldal egyik legnépszerűbb csoportja az „Antisemitizmust támogatók”. Az e csoportba belépők százalékos aránya határozottan korrelál az adott állam zsidó lakosságával. New York államban, ahol a legmagasabb a zsidók százalékos aránya, az „Antisemitizmust támogatók” csoport tagjainak számaránya is átlagon felüli.

Dna88 2001-ben csatlakozott a Stormfront-hoz. „Jóképű, fajilag tudatos”, harmincéves webfejlesztőként jellemzi magát, aki „Jew York Cityben” él. A belépését követő négy hónap során több mint kétszáz posztot írt, olyan tartalmakkal, mint: „A zsidók emberiségellenes bűnei”, „Zsidó vérdíj”, valamint a jewwatch.com weboldalra irányította az embereket, amely „a cionista bűnözés tudós könyvtáraként” hirdeti magát.

A Stormfront-tagok többnyire olyasmiket hánynak a kisebbségek szemére, hogy más nyelven beszélnek, illetve bűncselekményeket követnek el. De a magam részéről azt a panaszt találtam a legérdekesebbnek, hogy konkurenciát jelentenek „a randipiacon”.

Egy férfi, aki a William Lyon Mackenzie King nevet választotta magának (egy volt kanadai miniszterelnök után, aki egyszer felvetette, hogy „Kanadának a fehér ember országának kellene maradnia”), 2003-ban azt írta, nehe-

zen tudja „visszafogni magát”, amikor azt látja, hogy egy fehér nő „a ronda, korcs, félvér kishíggelével parádézik”. A Whitepride26 nevű nő pedig, egy 41 éves, Los Angeles-i diák, ezt írja a profiljában: „Utálom a feketéket, a latinokat, és olykor az ázsiaiakat is, különösen amikor a férfiak vonzóbbnak találják őket, mint egy fehér nőt.”

Bizonyos politikai fejlemények fontos szerepet játszanak. A Stormfront taglétszáma egyetlen nap alatt, egyetlen hullámban soha nem nőtt meg olyan kiugró mértékben, mint 2008. november 5-én, Barack Obama elnökké választásának másnapján. Akkor azonban egyáltalán nem nőtt a Stormfront iránti érdeklődés, amikor Donald Trump elnökjelölt lett, és akkor is csak egy kicsivel, amikor elnökké választották. Trump tehát meglovagolta ugyan a fehér nacionalizmust, de arra, hogy felkorbácsolta volna, nincs bizonyíték.

Obama megválasztása a fehér nacionalista mozgalom fellendülését eredményezte. Trump megválasztása erre adott reakciónak tűnik.

Ami viszont a jelek szerint nem számít, az a gazdasági helyzet. A havi tagregisztrációk száma és az adott állam munkanélküliségi rátája között egyáltalán nincs összefüggés. Azokban az államokban, amelyeket aránytalan mértékben sújtott a gazdasági válság, nem kerestek rá lényegesen többen a Stormfrontra.

De talán a legérdekesebbek – és egyúttal legmeglepőbbek – a Stormfront-tagok társalgási témái voltak. Nagyjából olyasmikről beszélgetnek, mint én a barátaimmal. Lehet, hogy ez a naivitásom jele, de úgy képzeltem, a fehér nacionalisták egy másik univerzumban élnek. Erre hosszú kommentfolyamokban méltatják a *Trónok harcát*, és megvitatják, hogy melyik online randioldal a jobb: a PlentyOfFish vagy az OkCupid.

A legfontosabb tény pedig, amely azt tanúsítja, hogy a Stormfront közönsége mégiscsak hasonló univerzumban él, mint a barátaim és én, a *The New York Times* népszerűsége. Nem csak VikingMaiden88 szokott a *Times* weboldalán lógni. Az oldal igencsak népszerű a tagság körében. Ami azt illeti, ha összehasonlítjuk a Stormfront-tagokat a Yahoo News látogatóival, kiderül, hogy az előbbieket kétszer olyan gyakran látogatnak el az nytimes.comra.

Egy gyűlöletoldal regisztrált tagjai a hű, de liberális nytimes.comot böngésszik? Ez meg hogy lehet? Ha a Stormfront-tagok jelentős hányada az nytimes.comról tájékozik, ez azt jelenti, hogy a fehér nacionalistákra vonatkozó közkeletű vélekedéseink tévesek. Ahogy az internet működésére vonatkozó közkeletű vélekedéseink is.

AZ IGAZSÁG AZ INTERNETRŐL

Szinte mindenki egyetért abban, hogy az internet eltávolítja egymástól az amerikaiakat, mivel az emberek többségét afelé tereli, hogy beássák magukat a hozzájuk hasonlóknak weboldalakba. Cass Sunstein, a Harvard Law School oktatója így írta le a helyzetet: „A kommunikációs piacunk gyors ütemben halad [egy olyan állapot felé, amikor] az emberek a saját nézőpontjukra korlátozzák magukat – a liberálisok zömmel vagy kizárólag liberálisokat néznek és olvasnak, a mérsékeltek mérsékelteket, a konzervatívok konzervatívokat, a neonácik neonácikat.”

Ez logikusan hangzik. Az internet gyakorlatilag korlátlan számú opciót kínál a hírfogyasztás terén. Azt olvasok, amit akarok. Ön is azt olvassa, amit akar. VikingMaiden88 is azt olvassa, amit akar. Márpedig az emberek, ha rájuk bízák a dolgot, jellemzően olyan nézőpontokat keresnek,

amelyek megerősítik azt, amiben eleve hisznek. Így az internet nyilvánvalóan a végletekig vitt politikai szegregáció melegegyháza.

Ezzel az alaptétellel csak egyetlen baj van. Az adatok tanúsága szerint egyszerűen nem igaz.

A közkeletű vélekedést cáfoló bizonyíték Matt Gentzkow és Jesse Shapiro 2011-es tanulmányából származik, két közgazdászéból, akikről korábban volt már szó.

A Gentzkow-Shapiro duó igen nagy elemszámú mintából gyűjtött adatokat az amerikaiak böngészési szokásairól. Az adathalmaznak részét képezte az alanyok – önbevalláson alapuló – ideológiai hovatartozása is: tudták róluk, hogy inkább liberálisnak, vagy inkább konzervatívnak tartják-e magukat. Az adataikat arra használták fel, hogy mérjék az interneten tapasztalható politikai szegregáció fokát.

Hogyan? Egy érdekes gondolat kísérlettel.

Tegyük fel, hogy találomra kiválasztunk két amerikaiat, akik történetesen ugyanazt a hírportált látogatják. Mekkorára a valószínűsége annak, hogy az egyik liberális, a másik pedig konzervatív? Más szóval: milyen gyakran „találkozik” a hírportálokon a két tábor?

Gondoljuk tovább a dolgot, és tegyük fel, hogy a liberálisok és a konzervatívok sosem tájékozódnak ugyanott az interneten. Vagyis a liberálisok kizárólag liberális weboldalakat látogatnak, a konzervatívok pedig kizárólag konzervatívokat. Ha így állna a dolog, akkor annak esélye, hogy két amerikainak, akik ugyanazon a hírportálon böngésznek, ellentétesek legyenek a politikai nézetei, 0 százaléka lenne. Az internet totálisan *szegregált* lenne: a liberálisok és a konzervatívok soha nem keverednének egymással.

Most pedig tegyük fel, hogy épp fordított a helyzet, vagyis a liberálisok és a konzervatívok között nincs semmi különbség abból a szempontból, hogy honnan tájékozód-

nak. Vagyis bármelyik hírportált ugyanolyan valószínűséggel keresik fel. Ha így állna a dolog, akkor annak esélye, hogy két amerikainak, akik ugyanazon a hírportálon böngésznek, ellentétesek legyenek a politikai nézetei, nagyjából 50 százalék lenne. Az internet totálisan *deszegregált* lenne: a liberálisok és a konzervatívok tökéletesen keverednének egymással.

Mit mondanak tehát Gentzkow és Shapiro adatai? Azt, hogy az Egyesült Államokban annak esélye, hogy két ember, akik ugyanazon a hírportálon böngésznek, különböző politikai nézeteket valljon, körülbelül 45 százalék. Az internet tehát sokkal közelebb áll a totális deszegregációhoz, mint a totális szegregációhoz: a liberálisok és a konzervatívok rendszeresen „találkoznak” egymással a neten.

Az internet szegregálatlanságát igazából az helyezi megfelelő kontextusba, ha összevetjük az életünk más területein tapasztalható elkülönültséggel. Gentzkow és Shapiro különféle offline interakciókkal is megismételte az elemzését. Mekkora az esély arra, hogy egy család két tagjának eltérő politikai nézetei legyenek? És két szomszédnak? Két kollégának? Két barátnak?

A General Social Survey adatai alapján arra jutottak, hogy mindezeknek kisebb az esélye, mint annak, hogy az ugyanazon hírportált látogató két embernek különbözzenek a politikai nézetei.

ANNAK VALÓSZÍNŰSÉGE, HOGY VALAKINEK,
AKIVEL TALÁLKOZUNK, MÁSOK A POLITIKAI NÉZETEI

Hírportálon	45,2%
Munkatárs	41,6%
Offline szomszéd	40,3
Családtag	37%
Barát	34,7%

Vagyis nagyobb valószínűséggel botlunk bele más nézeteket valló emberbe online, mint offline.

Hogyhogy nem szegregáltabb az internet? Azért, mert az online politikai szegregációt két tényező is korlátozza.

Az első talán kissé meglepően hangzik: az internetes híripart néhány nagy weboldal uralja. Általában hajlamosak vagyunk azt gondolni, hogy a világháló vonzza a szélsőségeket. És valóban mindenki talál kedvére való oldalakat, bármilyen nézeteket vall. Megvannak a gyülekezőhelyei a fegyvertartás mellett, illetve ellen keresztis hadjáratot folytatóknak, a szivarozók jogaiért vagy az egydolláros érmeért küzdő aktivistáknak, az anarchistáknak, a fehér nacionalistáknak. Az efféle webhelyek azonban csak kis töredékét teszik ki az internet hírforgalmának. 2009-ben a hírmegtekintések több mint felét négy oldal: a Yahoo News, az AOL News, az msnbc.com és a cnn.com gyűjtötte be. A Yahoo News továbbra is a legnépszerűbb hírportál az amerikaiak körében: havonta csaknem 90 millió egyedi látogatója van, ami mintegy 600-szorosa a Stormfront közönségének. A tömeges látogatottságú médiaoldalak, mint például a Yahoo News, széles és politikailag sokszínű közönséget vonzanak.

Az internet nem túl nagyfokú szegregáltságának másik oka az, hogy aki határozott politikai nézetekkel rendelkezik, az sokszor ellentétes véleményeket megjelenítő oldalakat is látogat, ha másért nem, hát azért, hogy felhergelje magát, és heves vitákat folytasson. A politikafüggők nem szorítkoznak a szájuk íze szerinti weboldalakra. Ha valaki látogatja a szélsőségesen liberális thinkprogress.orgot vagy a moveon.orgot, az az átlagos internethasználónál nagyobb valószínűséggel néz be a jobbra hajló foxnews.comra is. Ha valaki látogatja a szélsőségesen konzervatív rushlimbaugh.comot vagy a glennbeckot.comot, az az át-

lagos internethasználónál nagyobb valószínűséggel néz be a liberálisabb nytimes.comra is.

Gentzkow és Shapiro tanulmánya 2004 és 2009 közötti adatok alapján készült, ami az internet történetének viszonylag korai szakaszát fedi le. Lehetséges, hogy azóta sokkal inkább különálló rekeszekre tagolódtott a világháló? Felülírja-e Gentzkow és Shapiro eredményeit a közösségi média, különös tekintettel a Facebookra? Ha a barátaink jellemzően hasonló politikai nézeteket vallanak, mint mi, akkor a közösségi média elterjedése nyilvánvalóan a „visszhangkamrák” elszaporodását eredményezi. Igaz?

A történet megint csak nem ilyen egyszerű. Bár való igaz, hogy Facebook-barátaink elég nagy valószínűséggel osztják a politikai nézeteinket, egy adattudósokból álló kutatócsoport (név szerint Eytan Bakshy, Solomon Messing és Lada Adamic) rájött, hogy a Facebookon begyűjtött információinknak meglepően nagy része származik más nézeteiket valló emberektől.

Ez meg hogy lehet? Hát nem arról volt szó, hogy a barátaink általában ugyanazt gondolják a politikáról, mint mi? De igen, és így is van. Csakhogy van egy kulcsfontosságú tényező, amely miatt a Facebookon sokszínűbb politikai diskurzus zajlik, mint amikor offline társas életünket éljük. Az embereknek átlagban lényegesen több barátjuk van a Facebookon, mint offline. És ezek a Facebook által serkentett, lazább kötelékek nagyobb eséllyel jönnek létre ellentétes politikai nézeteket valló emberek között.

Vagyis a Facebook összehoz minket olyanokkal, akikkel kevésbé kötődünk: a középiskolai évfolyamtársunkkal, a bolond harmadunokatestvérünkkel és a barátunk barátjának a barátjával, akit mintha, talán, kábé ismer-nénk valahonnan. Velük aligha mennénk el bowlingozni

vagy kerti sütögetésre, aligha hívjuk meg őket vacsorára. De a Facebookon azért barátkozunk velük. És látjuk az általuk belinkelt cikkeket, amelyeket amúgy talán eszünkbe sem jutna elolvasni.

Összegezve: az internet valóban összehozza a különböző politikai nézeteket valló embereket. Könnyen lehet, hogy az átlagos liberális nő a liberális férjével és a liberális gyerekeivel tölti a reggeleit; majd a liberális kollégáival a délutánjait; és útközben is csupa liberális szlogent lát az autómaticákon; sőt még esténként, a jógacsoportban is csupa liberálissal mulatja az időt. A legnagyobb napi kitettségét a konzervatív hatásoknak az az időszak jelentheti, amikor hazaér, és átolvasgat néhány konzervatív kommentet a cnn.comon, vagy a republikánus középiskolai ismerőse belinkel valamit a Facebookon.

Valószínűleg én sem találkozom soha fehér nacionalistákkal a kedvenc brooklyni kávézómban. A *The New York Times* weboldalán viszont VikingMaiden88 ugyanolyan gyakran megfordul, mint én.

AZ IGAZSÁG A GYERMEKBÁNTALMAZÁSRÓL ÉS AZ ABORTUSZRÓL

Az internet nemcsak zavaró véleményekbe, hanem zavaró, sőt felkavaró viselkedésformákba is bepillantást nyújt. Olyannyira, hogy a Google-adatok egészen hatékonyan fel tudják hívni a figyelmet olyan krízisekre, amelyek a szokásos adatforrásokban nem jelennek meg. Az emberek akkor fordulnak a Google-hoz, ha baj van.

Nézzük meg például a gyermekbántalmazás alakulását a gazdasági válság idején!

Amikor 2007 végén kezdetét vette a recesszió, a szakértők magától értetődően aggódni kezdtek, hogy milyen hatással lesz mindez a gyerekekre. Nyilvánvaló, hogy ilyenkor sok a stresszes vagy depressziós tüneteket mutató szülő, ezek pedig komoly kockázati tényezők. Félő volt, hogy a gyermekbántalmazási ráta az egekbe szökhet.

Aztán befutottak a hivatalos adatok, és úgy tűnt, hogy az aggodalom alaptalan volt. A gyermekvédelmi szolgálatok jelentései arról számoltak be, hogy az ügyek száma csökkent. Sőt mi több, a válság által leginkább sújtott államokban csökkent a legnagyobb mértékben! „A világvégéjóslatok nem jöttek be” – nyilatkozta 2011-ben az *Associated Press*nek Richard Gelles, a Pennsylvaniai Egyetem gyermekjóléti szakértője. Igen, bármennyire kontraintuitívnak tűnik is, a gazdasági válság idején a gyermekbántalmazási ügyek száma meredeken zuhanni látszott.

De vajon tényleg kevesebb gyereket bántalmaztak a szóban forgó időszakban, amikor rengeteg felnőtt tengődött munka nélkül, nyomott hangulatban? Ezt nemigen tudtam elhinni. Ezért a Google-adatokat hívtam segítségül.

Mint kiderült, a gyerekek olykor szívszorító dolgokat írnak be a Google keresőjébe: olyanokat, hogy „anyukám megvert” vagy hogy „apukám megütött”. Az efféle keresések egészen más, meglehetősen tragikus képet festenek arról, ami akkoriban történt. A válság idején ugrásszerűen megnőtt az ezekhez hasonló keresések száma, szorosan követve a munkanélküliségi ráta alakulását.

Elmondom, mi történt szerintem: az, hogy csak a bejelentett esetek száma mutatott visszaesést, maga a gyermekbántalmazás azonban nem. A becslések szerint eleve csak az esetek kis hányadát jelentik be a hatóságoknak, és válság idején nagyobb az esély arra, hogy azok az emberek, akik

jelenteni szokták az ilyesmit (például tanárok, rendőrök), illetve akik foglalkoznak az ilyen ügyekkel (gyermekvédelmisek), túlterheltek, vagy éppen elvesztik állásukat.

Számos történet kering arról, hogy a válság idején azok, akik gyanús esetekről akartak bejelentést tenni, sokszor elunták a hosszú várakozást, és feladták a dolgot.

Tulajdonképpen van is bizonyíték arra (nem a Google jóvoltából), hogy a válság alatt gyakoribbá vált a gyermekbántalmazás. Ha egy gyerek bántalmazás vagy elhanyagoltság miatt életét veszti, azt mindenképpen jelenteni kell. Bár az ilyen halálesetek ritkák, a válság által leginkább sújtott államokban nőtt a számuk.

A Google pedig arra szolgáltat bizonyítékot, hogy a leginkább érintett térségekben többször gyanakodtak gyermekbántalmazásra az emberek. Ezekben a régiókban mind a válság előtti adatokhoz, mind az országos trendekhez képest nőtt az ilyen tárgyú keresések száma. A munkanélküliségi ráta minden egyes, egy százalékpontos növekedéséhez a „gyermekbántalmazás” és az „gyerek elhanyagolása” keresési rátájának három százalékos növekedése társult. A többségnek feltehetően nem sikerült jelentenie a bántalmazást, hiszen az ügyek száma épp ezekben az államokban csökkent a legnagyobb mértékben a hivatalos adatok szerint.

A szenvedő gyerekek által indított keresések száma nő. A halálesetek száma nő. A bántalmazásra gyanakvó emberek internetes kereséseinek száma nő. De a bejelentett esetek száma csökken. Úgy tűnik, a válság hatására több gyerek mondja el a Google-nak, hogy verik a szülei, és többen gyanakszanak arra, hogy bántalmazás tanúi. A túlterhelt hatóságok viszont kevesebb ügygel tudnak foglalkozni.

Azt hiszem, nagy biztonsággal kijelenthetjük, hogy a

gazdasági válság súlyosbította a gyermekbántalmazás problémáját, még akkor is, ha ez a hagyományos eszközökkel nem volt mérhető.

Amikor csak nem regisztrált emberi szenvedésre gyanakozom, mindig rögtön a Google-adatokhoz fordulok. Ennek az újfajta adatforrásnak – illetve annak, ha tudjuk értelmezni – az az egyik potenciális előnye, hogy segíthetünk olyan embereken, akik különben nem kerülnének a hatóságok látóterébe.

Ezért amikor a Legfelsőbb Bíróság nemrég az abortuszt szigorító törvények hatását vizsgálta, a keresési adatokat hívtam segítségül. Gyanítottam, hogy az új jogszabályok által érintett nők utánanézhettek a terhességmegszakítás nem hivatalos módjainak. Úgy is tettek, méghozzá legtöbbször az olyan államokban, amelyek az abortuszt korlátozó törvényeket hoztak.

A keresési adatok egyszerre hasznosak és nyugtalanítóak.

Az Egyesült Államokban 2015-ben több mint 700 ezer Google-keresést indítottak a magzatelhajtás témakörében. Összehasonlításképpen: az abortuszklinikákra abban az évben mintegy 3,4 millióan kerestek rá. Vagyis az abortuszt fontolgató nők jelentős hányadának megfordult a fejében, hogy saját maga szakítsa meg a terhességét.

Mintegy 160 ezerszer kerestek rá arra, hogyan lehet abortusztáblettához jutni nem hivatalos csatornákon: „abortusztábletta vásárlása online”, illetve „ingyenes abortusztábletta”. Faggatták a Google-t a zöldségek, például a petrezselyem és a C-vitamin által előidézett abortuszról is. Mintegy 4000-en kértek útmutatást a vállfával végzett abortusszal kapcsolatban, akik közül 1300-an szó szerint

ezt írták be: „hogyan kell vállfával abortuszt végezni”. Néhány százan pedig arra kerestek rá, hogyan kell úgy elhajtani a magzatot, hogy hipót juttatnak valaki méhébe, illetve hasba bokszolják az illetőt.

Vajon mi motiválja a magzatelhajtás iránti érdeklődést? A Google-keresések földrajzi és időbeli eloszlása valószínűleg leleplezi a felelőst. Az a helyzet, hogy ahol hivatalosan nehéz elvégeztetni az abortuszt, ott a nők utánanézőnek, hogyan lehetne elvégezni feketén.

A magzatelhajtásra irányuló keresések száma 2004 és 2007 között nagyjából stagnált. Csak 2008 végén, a pénzügyi válsággal és az azt követő recesszióval egyidejűleg kezdett emelkedni, majd 2011-ben nagyot ugrott: nem kevesebb mint 40 százalékot. A reprodukciós jogokkal foglalkozó Guttmacher Intézet 2011-et jelöli meg az országos szigorítási hullám kezdeteként. Abban az évben nem kevesebb mint 92 állami szintű jogszabály korlátozta az abortuszhoz való hozzáférést. Ugyanakkor Kanadában, ahol nem korlátozták a reprodukciós jogokat, a kérdéses időszakban nem nőtt ilyen mértékben a magzatelhajtásra irányuló Google-keresések száma.

Az ilyen keresések számaránya Mississippi államban a legmagasabb, ahol a hárommillió lakosságra egyetlen abortuszklinika jut. Azon tíz állam közül, ahol a legmagasabb a keresési ráta, nyolcat az abortusszal szemben „ellenségesnek” vagy „nagyon ellenségesnek” minősített a Guttmacher Intézet. Abból a tíz államból viszont, ahol a legalacsonyabb a keresési ráta, egyet sem soroltak ebbe a két kategóriába.

Azt persze nem tudjuk meg a Google-keresésekből, hogy hány nőnek sikerül a magzatelhajtás, de vannak olyan bizonyítékok, amelyek arra utalnak, hogy elég soknak. Az

igazság kiderítésének egyik módja az, hogy összevetjük az abortuszok számát a születési adatokkal.

2011 az utolsó év, amelynek teljes körű, állami szintű abortuszadatai rendelkezésre állnak. Ezen adatok szerint az olyan államokban, ahol kevés az abortuszklinika, sokkal kevesebb törvényes abortuszra került sor.

Most pedig hasonlítsuk össze azt a tíz államot, ahol a legmagasabb az egy főre jutó abortuszklinikák száma (köztük van például New York állam és Kalifornia), azzal a tíz állammal, ahol a legkevesebb abortuszklinika jut egy főre (utóbbi listán szerepel többek közt Mississippi és Oklahoma)! Az utóbbi tíz államban 54 százalékkal kevesebb törvényes abortuszt végeztek – azaz ezer, 15 és 44 év közötti nőre 11-gyel kevesebb abortusz jutott. Az élveszületések száma viszont magasabb volt a legkevesebb abortuszklinikával rendelkező államokban, de nem annyival, hogy önmagában magyarázza az abortuszok számában mutatkozó különbséget. Mindössze hattal több élveszületés jutott ezer szülőképes korú nőre.

Úgy tűnik tehát, hogy azokban az országrészekben, ahol a legnehezebb elvégeztetni az abortuszt, a terhességek egy része egyszerűen „eltűnt”. A hivatalos adatforrásokból nem derül ki, mi történt azzal az ezer nőre jutó öt terhességgel az abortuszt korlátozó államokban.

A Google viszont néhány eléggé használható nyomot biztosít. Nem bízhatunk meg vakon a kormányzati adatokban. A kormányzat kijelentheti, hogy csökkent a gyermekbántalmazási ügyek vagy éppen az abortuszok száma, és a politikusok ezt akár meg is ünnepelhetik, de könnyen lehet, hogy a szép eredmény egyedül a hibás adatgyűjtési módszernek tudható be. Könnyen lehet, hogy a valóság mást mutat, és az sem kizárt, hogy sokkal lehangelőbb.

AZ IGAZSÁG FACEBOOK-BARÁTAINKRÓL

Ez a könyv általánosságban a Big Datáról szól. A mostani fejezet azonban jórészt csak a Google-keresésekre helyezi a hangsúlyt, amelyek, amint azt már kifejtettem, rejtett világot tárnak fel előttünk, ez a rejtett világ pedig nagyon más, mint amit látni vélünk. Vajon más Big Data-források is digitális igazságszérumként szolgálnak? A helyzet az, hogy számos Big Data-forrás, így például a Facebook sokszor épp ellentétesen hat, mint a digitális igazságszérum.

A közösségi oldalakon, csakúgy, mint a felméréseken, semmi sem ösztönzi az embert arra, hogy igazat mondjon. A közösségi oldalakon, sokkal inkább, mint a felméréseken, erős ösztönzők hatnak az emberre, hogy jó színben tüntesse fel magát. Hiszen az efféle online jelenlét nem anonim jellegű. Az ember közönségnek udvarol, amikor elmondja barátainak, a rokonainak, a kollégáinak, az ismerőseinek, hogy ki is ő.

Hogy lássuk, mennyire torzítóan hathatnak olykor a közösségi médiából gyűjtött adatok, gondoljunk bele, mennyire népszerű a havonta megjelenő, tekintélyes, értelmiségieknek szóló *Atlantic* magazin, illetve a nemritkán szenzációhajhász, pletykafészek *National Enquirer* magazin. Mindkét kiadvány átlagos példányszáma nagyjából hasonló, néhány százezres. (A *National Enquirer* hetente jelenik meg, így az összpéldányszáma magasabb.) A két magazinra irányuló Google-keresések mennyiségére vonatkozóan is rendelkezünk összevethető számokkal.

A Facebookon azonban úgy másfél millióan vagy kedvelik az *Atlanticot*, vagy tárgyalják a cikkeit a profiljukon. Az *Enquirer*t mindeközben csupán mintegy 50 ezer ember kedveli, illetve ír róla.

AZ ATLANTIC, ILLETVE A NATIONAL ENQUIRER NÉPSZERŰSÉGE
KÜLÖNBÖZŐ FORRÁSOK SZERINT

Példányszám	Nagyjából azonos
Google-keresések száma	Nagyjából azonos
Facebook-kedvelések száma	27:1 az Atlantic javára

Ha egy magazin népszerűségét szeretnénk megbecsülni, ehhez a leghitelesebb alapot a példányszám biztosítja. A Google keresési adatai is ehhez közelítenek. A Facebook-adatok azonban erősen elfogultak a bulvárszenynlap kárára, ami által a Facebook a lehető legrosszabb adatforrás annak meghatározására, mit is kedvelnek valójában az emberek.

És nem csupán kedvelt olvasmányainkkal vagyunk így, hanem az egész életünkkel: a Facebookon a szalonképes énünket mutatjuk meg, nem pedig valódi önmagunkat. Használok a Facebook-adatokat ebben a könyvben, még ebben a fejezetben is, de mindig csak fenntartásokkal.

Hogy jobban megérthessük, mi mindenről nem tud még a közösségi média sem, kanyarodjunk vissza egy percre a pornóhoz! Először is arról az elterjedt hiedelemről kell szót ejtenünk, hogy az internetet előnti a szenny. Ez nem igaz: az online tartalmak nagy része nem pornográf. A tíz leglátogatottabb webhely között például egyetlen pornóoldal sincs. Vagyis a pornó népszerűségét, bár óriási, nem érdemes túlbecsülni.

Ezzel együtt is igaz, hogy ha közelebről szemügyre vesszük a nép pornókedvelési és -megosztási szokásait, egyértelműen kiderül: a Facebook, az Instagram és a Twitter csak korlátozott betekintést nyújt abba, hogy mi igazán népszerű a neten. A világhálónak hatalmas területei vannak, amelyek úgy örvendenek nagy népszerűségnek, hogy

a közösségi oldalakon csak minimális mértékben vannak jelen.

Egyetlen példa: minden idők legnépszerűbb videója e könyv írásakor a „Gangnam Style”. Psy, azaz Pak Cse-szang idétlen klipje, amely a trendi koreaiakat parodizálja, 2012-es megjelenése óta 2,3 milliárd megtekintést gyűjtött be a YouTube-on. De nemcsak ezen a felületen derül ki egyértelműen, hogy milyen népszerű, hanem szinte bárhol: több tízmilliószor osztották meg a közösségi média-platformokon.

Minden idők legnépszerűbb pornóvideója minden bizonnyal a „Great Body, Great Sex, Great Blowjob” („Szuper test, szuper szex, szuper szopás”). Ezt több mint 80 milliószor nézték meg, ami azt jelenti, hogy minden 30 „Gangnam Style”-megtekintésre jut egy „Great Body, Great Sex, Great Blowjob”-megtekintés. Ha a közösségi média pontos képet nyújtana a megtekintett videókról, a „Great Body, Great Sex, Great Blowjob”-nak több millió megosztást kellett volna produkálnia. De mindössze néhány tucatnyian osztották meg, ráadásul minden alkalommal pornósztrók, és sohasem átlagos felhasználók. Az emberek magától értetődően nem éreznek késztetést arra, hogy a pornó iránti érdeklődésüket a barátaik előtt reklámozzák.

A Facebook nem digitális igazságszérum, hanem amolyan „hencegünk a hiper-szuper életünkkel”-szérum. A Facebook világában a jelek szerint az átlagos felnőtt boldog házasságban él, a Karib-tengernél vakációzik, és az *Atlanticot* böngészi. A való világban sokan kukkantanak bele a *National Enquirer*-be, miközben dűhtől tajtékozva sorban állnak a szupermarket pénztáránál, és nem veszik fel a telefont, amikor a hitvesük hívja őket, akivel évek óta nem hálnak együtt. A Facebook világában a családi élet tökéletesnek tűnik. A való világban a családi élet zűrös. Néha

annyira zűrös tud lenni, hogy egyesek azt is megbánják, hogy gyereket vállaltak. A Facebook világában a jelek szerint szombat este minden fiatal felnőtt valami menő buliban tombol. A való világban a többség otthon ül, és orrvérzésig nézi a sorozatokat a Netflixen. A Facebook világában a lányok huszonhat, boldogságban fürdőző képet posztolnak arról, hogy kiruccantak valahová a barátjukkal. A való világban a lány közvetlenül a fotók posztolása után beírja a Google keresőjébe, hogy „a barátom nem akar szexelni velem”. A barátja pedig talán pont ugyanakkor a „Great Body, Great Sex, Great Blowjob”-ot nézi.

DIGITÁLIS IGAZSÁG	DIGITÁLIS HAZUGSÁGOK
Keresések	Posztok a közösségi oldalakon
Megtekintések	Kedvelések a közösségi oldalakon
Kattintások	Randiprofilok
Online vásárlások	

AZ IGAZSÁG ÜGYFELEINKRŐL

A Facebook 2006. szeptember 5-én kora reggel jelentős frissítést hajtott végre a honlapján. A korábbi verziók még csak azt tették lehetővé, hogy a felhasználók rákattintsanak a barátaik profiljára, és megnézzék, mit csinálnak. A nagy sikerként számon tartott weboldalnak akkor 9,4 millió felhasználója volt.

Aztán a Facebook mérnökei több havi kemény munka árán létrehoztak egy általuk „News Feed”-nek, azaz hírfolyamnak nevezett dolgot, amely arra szolgált, hogy a felhasználók frissítéseket kapjanak az összes barátjuk tevékenységéről.

A felhasználók rögtön kijelentették, hogy gyűlölik a hír-

folyamot, és Ben Parr, a Northwestern Egyetem hallgatója létrehozta a „Students Against Facebook news feed” („Diákok a Facebook-hírfolyam ellen”) nevű csoportot. Parr így nyilatkozott: „A hírfolyam egyszerűen ijesztő: olyan, mint ha suttyomban követnél valakit. Ennek a dolognak el kell tűnnie!” A csoportnak alig néhány nap alatt 700 ezer tagja lett, akik osztották Parr érzéseit. A Michigani Egyetem egyik harmadévese szinte szóról szóra ugyanazt mondta a *Michigan Daily*nek, mint ő: „Megrémít az új Facebook. Olyan, mintha valakinek a nyomában koslatnék.”

David Kirkpatrick *A Facebook-jelenség. Egy bennfentes sztorija a cégről, amely összeköti a világot* című könyvében, amely a weboldal történetét dolgozza fel, ezt a sztorit is felidézi. A hírfolyam bevezetését a legsúlyosabb válságnak minősíti, amellyel a Facebooknak valaha szembe kellett néznie. De arról is tudósít, hogy amikor meginterjúvolta Mark Zuckerberget, a gyorsan növekvő cég társalapítóját és vezetőjét, azt egyáltalán nem hatotta meg a tiltakozás.

Hogy miért? Azért, mert Zuckerberg hozzáfért a digitális igazságszérumhoz: az emberek Facebook-látogatásainak, illetve Facebook-kattintásainak számához. Kirkpatrick így ír erről:

Zuckerberg igazából tudta, hogy az embereknek tetszett a hírek, bármit is írtak a csoportokban. Az adatai ezt igazolták: a felhasználók átlagosan több időt töltöttek a Facebookon, mint mielőtt a hírek elindult. És több dolgot csináltak, még hozzá jóval többet: augusztusban a felhasználók 12 milliárd oldalt néztek meg, de októberben, a hírek elindulásával 22 milliárdot.*

* David Kirkpatrick: *A Facebook-jelenség*. Budapest, Gabo, 2011, 211. Czibik Márta fordítása. (A szerk.)

És nem ez volt az egyetlen adat, amely bizonyította Zuckerberg igazát. A hírfolyamellenes csoport népszerűsége is arról tanúskodott, hogy a hírfolyam hatalmas erő! A csoport azért tudott olyan gyorsan növekedni, mert rengetegen hallották, hogy a barátaik beléptek – és azért szereztek erről tudomást, mert volt hírfolyamuk.

Vagyis bekapcsolódtak az össznépi felhördülésbe, de miközben azon háborogtak, hogy milyen szörnyű a barátaik életének minden rezdülését látni a Facebookon, mindig visszatértek a Facebookra, hogy buzgón figyeljék a barátaik életének minden rezdülését. A hírfolyam maradt, a Facebooknak pedig ma már több mint egymilliárd napi szintű aktív felhasználója van.

Peter Thiel, aki a kezdeti időszakban pénzt fektetett a Facebookba, *Nulláról az egyre* című könyvében azt írja, hogy a nagy vállalkozások mindig titkokra épülnek: vagy a természet, vagy az emberek titkaira. Amint azt a 3. fejezetben már láttuk, Jeff Seder rájött a természet azon titkára, hogy a versenyllovak jövőbeni teljesítményét jól megjósolja a bal szívkamra mérete. A Google pedig rájött a természet azon titkára, hogy milyen elképesztő hatalmat jelent a linkekben rejlő információ.

„Az emberek titkait” Thiel így definiálja: „olyan dolgok, amiket az emberek nem tudnak magukról vagy rejtegetnek, mert nem akarják, hogy mások megtudják”.* Más szóval: az ilyen vállalkozások az emberek hazugságaira épülnek.

Akár amellett is érvelhetnénk, hogy az egész Facebook alapja egy kínos titok, amelyről a cég alapítója a Harvardon szerzett tudomást. Zuckerberg másodév elején létre-

* Peter Thiel és Blake Masters: *Nulláról az egyre. Gondolatok a startupokról, avagy hogyan teremtsünk jövőt*. Gabo, 2015, 108. Darnyik Judit fordítása. (A szerk.)

hozott a diáktársai számára egy Facemash nevű honlapot. Az „Am I Hot or Not?” („Dögös vagyok vagy sem?”) nevű weboldaltól mintázott Facemashen két harvardos diák közül kellett választani, hogy melyikük néz ki jobban.

A másodéves fiú ötlete felháborodást váltott ki. Az egyetemi újság, a *Harvard Crimson* vezércikkében szemére hányta az ifjú Zuckerbergnek, hogy „kiszolgálja az emberek legsötétebb oldalát”. Spanyolajkú és afroamerikai csoportok pedig szexizmussal és rasszizmussal vádolták. De mire az egyetem adminisztrátorai törölték Zuckerberg nettozáférését – vagyis mindössze néhány óra leforgása alatt –, már 450-en tekintették meg a Facemash-t, és összesen 22 ezer szavazat érkezett a különböző képekre.

Zuckerberg megtudott egy fontos titkot: mondhatják az emberek, hogy fel vannak háborodva, ízléstelennek is bélyegezhetik, de attól még rá fognak kattintani.

És megtudott még valamit: azt, hogy azok is szívesen osztályozzák a társaik külsejét, akik amúgy komoly, felelősségteljes, mások privát szféráját tiszteletben tartó embernek vallják magukat – lám, még a Harvard diákjai is! Erről árulkodtak a megtekintések és a szavazatok. Később pedig, mivel a Facemash sok vitát váltott ki, Zuckerberg másként kamatoztatta frissen szerzett tudását. Rájött, hogy az emberek szívesen gyűjtenek felszínes információkat olyanokról, akiket félig-meddig ismernek, és erre alapozva létrehozta nemzedéke legsikeresebb vállalatát.

A Netflix is ehhez hasonló tanulságot szűrt le működése kezdeti szakaszában: ne higgy annak, amit az emberek mondanak neked; annak higgy, amit tesznek!

A cég eredetileg lehetővé tette, hogy a felhasználók sorrendbe állítsák azokat a filmeket, amelyeket valamikor megnéznének, de most nincs rá idejük. Így később emlékeztetni tudta őket a kiválasztott filmjeikre.

Csak hogy a Netflix figyelte egy furcsaságra az adatokban. A felhasználók összeírtak egy rakás filmet, de amikor néhány nappal később megkapták az emlékeztetőt, csak ritkán kattintottak rá valamelyikre.

Vajon mi lehetett a probléma? Hát az, hogy ha az embereket megkérdezzük, milyen filmeket néznének meg néhány napon belül, csupa igényes, intellektuális igényű alkotást sorolnak fel, például fekete-fehér dokumentumfilmeket a második világháborúról és külföldi művészfilmeket. Néhány napot később viszont már szívesebben nézik azt, amit szoktak: alpári vígjátékokat és romantikus limonádékat. Vagyis következetesen hazudnak önmaguknak.

Amikor a Netflix szembesült ezzel az ellentmondással, felhagyott azzal, hogy arról kérdezzesse az embereket, mit néznének meg a közeljövőben, és helyett hozzálatott egy olyan modell megalkotásához, amely a filmfogyasztók millióinak kattintásain és megtekintésein alapult. A cég attól kezdve nem olyan filmeket javasolt megtekintésre, amelyekről a felhasználók azt mondták, hogy szeretik őket, hanem az adatok alapján állította össze a listát. Az eredmény: a felhasználók gyakrabban látogattak el a Netflixre, és több filmet néztek meg.

„Az algoritmusok jobban ismernek téged, mint te magadat” – mondja Xavier Amatriain, a Netflix volt adattudósa.

MIÉRT OLYAN ESZMÉLETLEN JÓ, HA FÜTYÜLÜNK RÁ, MIT MONDANAK AZ EMBEREK?

AMIT MONDANAK	A VALÓSÁG	IPSO FACTO...
Nem szeretnék kémkedni a barátaim után.	Semmire sem vágnak jobban, mint hogy mindent tudjanak a barátaikról, és ítélkezhesnek felettük.	Mark Zuckerberg, a Facebook társalapítója 55,2 milliárd dolláros vagyont halmozott fel.

AMIT MONDANAK	A VALÓSÁG	IPSO FACTO...
Nem szeretnék munkásnyűző üzembengyártott termékeket vásárolni.	„Jutányos árú” termékeket vásárolnak.	Phil Knight, a Nike társalapítója 25,4 milliárd dolláros vagyont halmozott fel.
Szeretnék reggelente híreket hallgatni.	Szeretnék reggelente pornósztorokkal szexelő törpékről hallani.	Howard Stern 500 millió dolláros vagyont halmozott fel.
Nem érdekli őket a kikötőzős szex, a dominancia és a szado-mazochizmus.	Szeretnék az egyetemista lány és a dúsgazdag üzletember közti BDSM-szextről olvasni.	A szürke ötven árnyalatából 125 millió példány fogyott el.
Szeretnék, hogy a politikusok körvonalazzák a konkrét szakpolitikai elképzeléseiket.	Szeretnék, hogy a politikusok megkíméljék őket a részletektől, de keménynek és magabiztosnak tűnjenek.	Donald Trump

EL TUDJUK-E VISELNI AZ IGAZSÁGOT?

A fejezet egyes részeit alkalmasint csüggesztőnek találhatják. A digitális igazságszérum a következőkre derített fényt: makacsul tartja magát az a késztetés, hogy a külsőjük alapján ítéljük meg az embereket; még mindig több millió meglepő bujkál; a nők nem elhanyagolható hányada nemi erőszakról ábrándozik; széles körben elterjedt az afroamerikaiakkal szembeni gyűlölet; válságos a helyzet a gyermekbántalmazás és a magzatelhajtás terén; a dühödtt iszlamofóbia csak fokozódott, amikor az elnök toleranciára szólított fel. Nem éppen örömteli dolgok. Amikor előadást tartok valahol a kutatásaimról, az emberek sok-

szor odajönnek hozzám a végén, és azt mondják: „Seth, ez mind nagyon érdekes. De *annyira* nyomasztó!”

Nem próbálok úgy tenni, mintha egyes adatok nem lennének valóban kissé vészjóslóak. Ha az embernek egyfolytában azt mondják, amiről azt gondolják, hogy hallani akar, akkor általában olyasmit fog hallani, ami sokkal vigasztalóbb, mint az igazság. Ezért a digitális igazságszérum jellemzően arra világít rá, hogy a világ rosszabb, mint gondoltuk.

De tényleg hiányzik ez nekünk? Mindazok után, amiket megtudtunk a Google-keresésekről, a pornóadatokról és arról, hogy ki mire kattintgat, valószínűleg nem azt fogjuk gondolni, hogy: „Nagyszerű! Végre tudjuk, milyenek vagyunk valójában!” Hanem ezt: „Ez borzasztó! De legalább végre tudjuk, milyenek vagyunk valójában.”

Az igazság mégis segít – és nem is csak Mark Zuckerbergnek és a többi kattintásvadásznak. Ugyanis legalább háromféleképpen javíthat az életminőségünkön ez a tudás.

Először is vigaszt nyújthat, hogy nem vagyunk egyedül: nem csak minket gyötörnek kételyek, és nem csak mi szoktunk kínosan viselkedni. Jó lehet tudni róla, hogy mások is bizonytalanok a testüket illetően. Valószínű, hogy sokaknak – különösen azoknak, akik nem szexelnek túl gyakran – jó lehet tudni róla, hogy nem mindenki bagzik egyfolytában, mint a nyulak. Egy Mississippi államban élő középiskolás fiú számára pedig, aki bele van zúgva az amerikaifoci-csapat irányítójába, kifejezetten értékes információ lehet, hogy ha kevés is a környezetében a nyíltan melegférfi, rajta kívül még sokan érezhetnek efféle vonzalmat.

Az életnek akad még egy területe – erről eddig még nem esett szó –, ahol a Google-keresések szintén megnyugtathatnak minket, hogy nem vagyunk egyedül. Gyerekkor-

runkban bizonyára sokszor hallottuk tanárainktól, hogy akinek kérdése van, bátran tegye fel a kezét, mert ha nem ért valamit, akkor nyilván nincs ezzel egyedül. Na most, aki egy kicsit is hasonlít rám, az fittyet hányt a tanár biztatására, csak üldögélt tovább, és ki se merté nyitni a száját. Ő is arra gondolt, hogy biztosan hülyeséget kérdezne – bezzeg mások csupa okos, mélyenszántó kérdést tesznek fel. A névtelen, összesített Google-adatok viszont egyszer s mindenkorra meggyőznek minket arról, hogy mennyire igazuk volt a tanárainknak. Mások fejében is rengeteg felület, távolról sem mélyenszántó kérdés rejtőzik.

Vegyük például azokat a kérdéseket, amelyek Obama 2014-es évértékelő beszéde alatt a leginkább foglalkoztatták az amerikaiakat! (A színes fotót lásd a könyv végén.)

NEM CSAK NEKED JUTOTT ESZEDBE: A LEGGYAKORIBB
GOOGLE-KÉRDÉSEK AZ ÉVÉRTÉKELŐ BESZÉD ALATT

Hány éves Obama?
Ki ül Biden mellett?
Miért van Boehneren zöld nyakkendő?
Boehner miért narancssárga?

Olvasás közben most talán arra gondolnak, milyen gyászos képet festenek ezek a kérdések a demokráciánk állapotáról. Elég rosszul veszi ki magát, hogy az elnöki beszéd tartalma helyett valakinek a nyakkendőjével meg az arcszínével vagyunk elfoglalva. Ahogy az sem árul el túl sok jót a politikai aktivitásunkról, ha nem ismerjük fel John Boehner-t, aki akkor a Kongresszus képviselőházi elnöke volt.

Én viszont inkább úgy fogom fel a dolgot, hogy az efféle kérdések egykori tanáraink bölcsességét igazolják. Olyan kérdések ezek, amelyeket az emberek rendszerint nem

tesznek fel, nehogy hülyének nézzék őket. De akkor is sokkalban felmerülnek – és ha már felmerülnek, be is írják azokat a Google-ba.

Valójában úgy vélem, hogy a Big Data amolyan huszonegyedik századi változata lehet a híres önségítő mantrának, amely így szól: „Soha ne hasonlítsd a bensődet a többi ember külsejéhez!”

A Big Datá-s változat ez lehetne: „Soha ne hasonlítsd a Google-kereséseidet a többi ember közösségi oldalakon közzétett posztjaihoz!”

Hasonlítsuk össze inkább azt, hogy miként jellemzik a nők a férjüket a nyilvános közösségi oldalakon, illetve a névtelen keresésekben.

FÉRJEK LEGGYAKORIBB JELZŐI

KÖZÖSSÉGI MÉDIA	KERESÉSEK
a legjobb	meleg
a legjobb barátom	egy gyökér
hihetetlen arc	hihetetlen arc
a legszuperebb	idegesítő
„olylan cukil!”	aljas

Mivel mások posztjait látjuk, a kereséseiket viszont nem, hajlamosak vagyunk eltúlozni azoknak a nőknek a számát, akik egyfolytában „a legjobbnak”, „a legszuperebbnek” és „olylan cukinak” tartják a férjüket.* Hajlamosak vagyunk

* Twitter-adatokat elemeztem. Köszönettel tartozom Emma Piersonnak, amiért segített letölteni az adathalmazt. Nem vettem figyelembe az arról való beszámolókat, hogy épp mit csinál a férj az adott pillanatban; igaz, hogy a közösségi média tele van ilyen leírásokkal, keresőkifejezéseként azonban az ilyesmi nem igazán értelmezhető. Még ezeknek a tevékenységeknek a leírása is valószínűleg jó színben tüntet fel a férjeket. A leggyakrabban említett két tevékenység az, hogy a férj „dolgozik”, illetve „főz”.

alábecsülni azon nők számát, akik szerint a férjük „egy gyökér”, „idegesítő” vagy „aljas”. A névtelen, összesített adatok elemzése megértetheti velünk, hogy nem csak számunkra jelent kemény diót a házasság és az élet. Talán le szokhatunk arról, hogy mások posztjaihoz hasonlítgassuk a kereséseinket.

A digitális igazságszérum másik előnye, hogy figyelmeztethet minket a másik ember szenvedésére. A Human Rights Campaign nemrég kért meg arra, hogy segítsek nekik abban a programjukban, amelynek célja, hogy bizonyos államokban felvilágosítsák a melegférfiakat az előbújás lehetőségéről. Azt szeretnék eldönteni a névtelen, összesített Google-keresések segítségével, hogy ennek során hova érdemes összpontosítani az erőforrásaikat. A gyermekvédelmi hatóságok is felvették velem a kapcsolatot: arra kíváncsiak, hogy az ország mely részén fordulhat elő több gyermekbántalmazási ügy, mint amennyit rögzítenek.

Egy meglepő témában is érkezett megkeresés: a hüvely-szag ügyében. Amikor először írtam a témáról a *The New York Times*-ban – hol máshol? –, ironikus hangot ütöttem meg. Másokkal együtt én is jót kuncogtam ezen a részen.

Mindazonáltal amikor később elmélyedtem kicsit azokban az üzenőfalakban, amelyek a keresés nyomán előjöttek, számos posztba botlottam olyan kislányoktól, akik arra panaszkodtak, hogy tönkremegy az életük, mert folyton a nemi szervük szaga miatt görcsölnék. Ami már egyáltalán nem vicces. Szexuális felvilágosítással foglalkozó szakemberek érdeklődtek nálam, hogy miként vehetnék hasznát az internetes adatoknak, ha vissza akarják szorítani a kislányok körében terjedő paranoiát.

Ilyenkor egy kicsit mindig úgy érzem, hogy kevés vagyok én az ilyesmihez. Mindazonáltal kétségtelenül ko-

moly dolgokról van szó, és meg vagyok győződve róla, hogy az adattudomány a segítségünkre lehet.

A digitális igazságszérum legutolsó – és szerintem legfőbb – értéke valójában éppen abban rejlik, hogy képes elvezetni bennünket a problémától a megoldásig. Ha sikerül többet megértenünk a világból, akkor talán módot találhatunk arra, hogy eredménnyel vehessük fel a harcot a világot fertőző ocsmány előítéletek szinte kimeríthetetlen sokaságával.

Térjünk csak vissza Obama iszlamofóbiáról szóló beszédéhez! Idézzük fel, mi is történt: az elnök amellett próbált érvelni, hogy tanúsítsunk több tiszteletet a muszlimok iránt; az emberek azonban, akikhez szólni próbált, csak még jobban feldühödtek.

A Google-keresésekből azonban kiderül, hogy a beszéd egyik mondatának mégis sikerült olyan reakciót kiváltania, amely talán megfelelt az akkori elnök szándékának. A szóban forgó mondat a következő volt: „A muszlim amerikaiak a barátaink, a szomszédaink, a munkatársaink, bálványozott élsportolóink, és igen, vannak köztük egyenruhás férfiak és nők is, akik, ha kell, az életüket áldozzák országunk védelmében.”

Miután ez a mondat elhangzott, a „muszlim” kezdetű Google-keresésekben – több mint egy év óta először – nem a „terroristák”, a „szélsőségesek” vagy a „menekültek” volt a leggyakoribb folytatás, hanem a „sportoló” és a „katona”. Sőt mi több, a „sportoló” egy teljes napon át az élen maradt.

A keresési adatok arra utalnak, hogy ha a dühös embereket kioktatjuk, azzal csak még jobban felbőszítjük őket. Ha viszont finoman felcsigázzuk a kíváncsiságukat, ha új információkkal szolgálunk nekik, ha az addigitól eltérő képet festünk az indulataikat kiváltó csoportról, akkor talán pozitívabb irányba tudjuk terelni a gondolataikat.

Obama két hónappal a San Bernardinó-i mérsárlás után újabb televíziós beszédet mondott az iszlamofóbiáról, ezúttal egy mecsetben. Az elnök stábjából talán olvasta valaki a *The New York Times*, ahol Evan Soltasszal megírtuk a rovatunkban, hogy mi működött, és mi nem. A beszéd tartalma ugyanis érezhetően más volt, mint az előzőé.

Obama ezúttal nem helyezett különösebb hangsúlyt arra, hogy milyen nagy érték a tolerancia. Inkább arra összpontosított, hogy felkeltse az emberek érdeklődését, és megváltoztassa a muszlim amerikaiakra vonatkozó percepcióikat. Elmondta például, hogy az Afrikából behurcolt rabszolgák közül sokan muszlimok voltak; elmondta, hogy Thomas Jeffersonnak és John Adamsnek is volt saját példánya a Koránból; elmondta, hogy amerikai földön Észak-Dakotában épült az első mecset; elmondta, hogy volt egy muszlim amerikai, aki felhőkarcolókat tervezett Chicagóban. Ismét említést tett a muszlim sportolókról és a fegyveres erők tagjairól, de beszélt muszlim rendőrtisztekről, tűzoltókról, tanárokról, orvosokról is.

A Google-kereséseket elemezve megállapítottam, hogy ez a beszéd alighanem nagyobb sikert ért el, mint az előző. Az elhangzása utáni órákban csökkent a muszlimokkal szembeni gyűlöletet, indulatokat tükröző keresések száma.

Más módja is van annak, hogy a keresési adatok segítségével kiderítsük, mi okozza, illetve mi csökkenti a gyűlöletet. Megnézhetjük például, hogyan változik a rasszista töltetű keresések száma, miután egy város amerikaifutball-csapata fekete irányítót szerződtet, vagy a szexista keresések száma, miután egy nőt megválasztanak valamilyen komoly posztra. Megnézhetjük azt is, hogyan hat a rasszizmusra a közösségi rendészet, vagy hogy hogyan hatnak a szexizmusra az új szexuális zaklatási törvények.

Hasznos lehet az is, ha tudomást szerzünk a saját, tudat alatt működő előítéleteinkről. Törekedhetünk például arra, hogy többet gyönyörködjünk a kislányok szellemi képességeiben, és kevesebbet aggodalmaskodjunk a külsejük miatt. A Google keresési adatai és az igazság más internetes kútforrásai példátlan mélységű betekintést engednek az emberi elme legsötétebb zugaiba. Belátom, hogy ezekbe a sötét zugokba időnként nehéz bekukkantani. De mind-ebből erőt meríthetünk. Az adatokat felhasználva felvehetjük a harcot a sötétséggel. Azzal, hogy rengeteg adatot összegyűjtünk a világ problémáiról, megteesszük az első lépést a leküzdésük felé.

5. fejezet

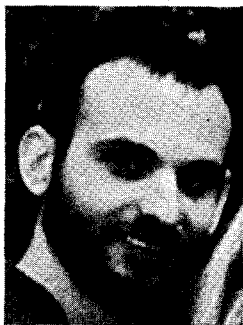
Zoomoljunk rá!

Az öcsém, Noah négy évvel fiatalabb nálam. Akik először látnak minket, többnyire úgy vélik, hogy kísértetiesen hasonlítunk egymásra. Mindketten jó hangosan beszélünk, egyformán kopaszodunk, és meggyűlik a bajunk a lakásunk rendben tartásával.

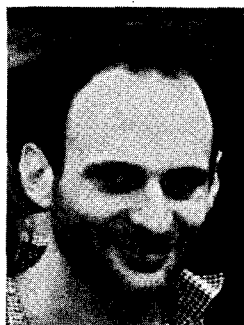
De azért akadnak különbségek is. Én a fogamhoz verem a garast, Noah viszont mindenből a legjobbat veszi meg. Én odavagyok Leonard Cohenért és Bob Dylanért, Noah-nak viszont a Cake jön be és Beck.

A legszembeötlőbb különbség talán mégis a baseball-hoz fűződő viszonyunkban fedezhető fel. Én megszállottja vagyok ennek a sportnak, a New York Mets imádata pedig különösképpen az identitásom egyik fő alapját képezi, amióta csak az eszemet tudom. Noah ezzel szemben végtelesen unalmasnak tartja a baseballt, és a sportág gyűlölete már régóta az identitásának az alapja.*

* A teljes átláthatóság jegyében hozzá kell tennem, a tényellenőrzés során Noah tagadta, hogy e népszerű amerikai kedvtelés gyűlölete meghatározóná a személyiségét. Azt beismeri, hogy utálja a baseballt, de hite szerint a személyisége „magját” a kedvessége, a gyerekszeretete és az intelligenciája alkotja. Úgy gondolja, hogy a baseballhoz való viszonya „az első tízbe” sem kerülne be. De én azt a következtetést vontam le, hogy az ember a saját identitását sokszor nem látja objektíven. Külső megfigyelőként viszont láthatom, hogy Noah-nak alapvető jellemvonása a baseballellenesség, akár felismeri ezt ő maga, akár nem. Ezért benne maradt a könyvben.



Seth Stephens-Davidowitz,
baseballofil



Noah Stephens-Davidowitz,
baseballfófb

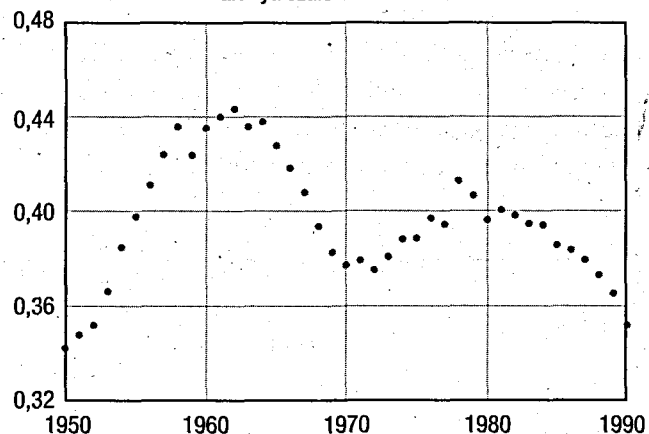
Hogy lehet az, hogy két, ennyire hasonló génállományú fickó, akiket ugyanazok a szülők neveltek fel ugyanabban a városban, ennyire eltérően viszonyul a baseballhoz? Mi határozza meg, hogy milyenek leszünk felnőttként? És felmerül egy még alapvetőbb kérdés: mi lehet a *baj* Noah-val? Van a fejlődéslélektanon belül egy egyre fejlődő szakterület, amely gigantikus méretű felnőtt adatbázisokból bányász ki adatokat, és összepárosítja azokat a gyermekkor kulcsfontosságú eseményeivel. Ez a tudományterület segítségünkre lehet abban, hogy megbirkózzunk az ehhez hasonló kérdésekkel. Akár Big Psych-nak, azaz „nagy pszichónak” is elnevezhetjük, hiszen a „nagy adat” pszichológiai kutatásokhoz való, egyre intenzívebb felhasználásán alapul.

A „Big Psych” működését egy saját készítésű tanulmányral illusztrálnám, amely arról szól, hogy miként befolyásolják gyermekkori élményeink a baseballcsapat-választásunkat, illetve azt, hogy baseballszurkoló lesz-e belőlünk egyáltalán. E célból baseballcsapatok lájkjaira vonatkozó Facebook-adatokat használtam fel. (Az előző fejezetben kitértem arra, hogy a Facebook-adatok súlyosan félre-

zetőek tudnak lenni, ha érzékeny témáról van szó. Most viszont abból indultam ki, hogy egy baseballcsapat iránti rajongását senki sem szégyelli bevallani a Facebookon. Talán még a Phillies-drukkerek sem.)

Először is letöltöttem azoknak a különböző életkorú férfiaknak a számát, akik lájkolták a két New York-i baseballcsapat valamelyikét. Íme, a Mets-szurkolók százalékos aránya születési év szerinti bontásban:

A Metset lájkoló New York-i férfi baseballdrukkerek százalékos aránya születési év szerint

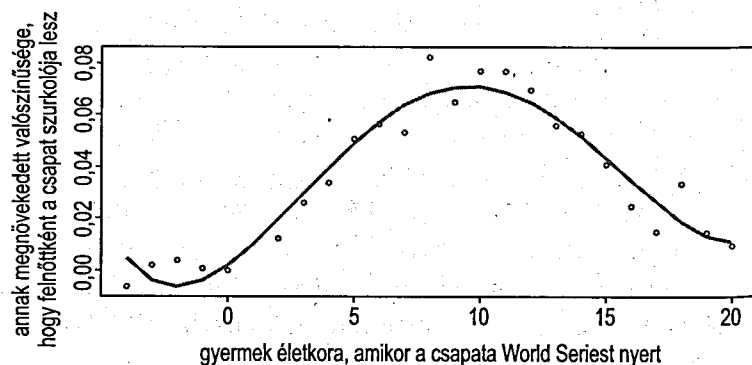


Minél magasabban jár a grafikonon a pötty, annál több az adott „évjáratban” a Mets-drukker. Mint látható, a csapat népszerűsége nő, majd csökken, majd megint nő, és megint csökken. Az is látható, hogy a Mets roppant népszerű az 1962-es, illetve 1978-as születésűek körében. Azt hiszem, a baseballdrukkerek már kapiskálják, hogy mi az ábra. Arról van szó, hogy a Mets csupán kétszer nyerte meg a World Seriest, vagyis a bajnoki döntőt: 1969-ben és 1986-ban. Ezek a férfiak durván hét-nyolc éves kisráccok voltak,

amikor a Mets bajnok lett. A Mets-szurkolóvá válást tehát, legalábbis a fiúk esetében, nagy pontossággal előre jelzi, ha a Mets hét-nyolc éves korukban bajnokságot nyer.

Ami azt illeti, ki is terjeszthetjük az elemzés hatósugarát. Meg is tettem: letöltöttem a Facebook-információkat arra vonatkozóan, hogy hány, különböző életkorú rajongó lájkolja a Major League Baseball különböző csapatait.

Kiderült, hogy a Baltimore Oriolest lájkoló férfiak között szokatlanul sokan születtek 1962-ben, míg a Pittsburgh Pirates-rajongók közül 1963-ban. Ezek az emberek mind hét-nyolc évesek voltak akkor, amikor ezek a csapatok megnyerték a World Seriést. Miután az összes vizsgált klub esetében kiszámoltam, hogy melyik évjáratból való a legtöbb rajongója, illetve hány évesek lehettek, amikor a csapatuk bajnok lett, a következő grafikont kaptam:



Megint csak azt látjuk, hogy egy férfi életében nyolcéves kora körül érkezik el a legfontosabb év abból a szempontból, hogy felnőttként melyik lesz a kedvenc baseballcsapata. Az a tágabb időszak pedig, amikor egy fiút a leginkább meg lehet nyerni az ügynek, az 5–15 éves kor. A 19–20 éveseknél már csak nyolcadannyi jelentősége van a szurkoló-

vá válás szempontjából annak, hogy egy bizonyos csapat World Seriést nyer-e. Addigra ugyanis általában eldől a kérdés: az ember vagy összeköti magát egy életre valamelyik csapattal, vagy nem.

Joggal szegeznék most nekem a kérdést, hogy mi a helyzet a baseballrajongó nőkkel. Náluk nem rajzolódik ki ilyen határozott mintázat, de úgy tűnik, hogy a kulcsfontosságú életkor az ő esetükben 22 év.

Ez a kedvenc tanulmányom; hiszen két hón szeretett témámról is szó esik benne: a baseballról, illetve felnőttkori bosszúságaim forrásáról. 1986-ban egyszer és mindenkorra horogra akadtam, és azóta is viselem a Mets-drukker-ség minden kínját. Noah-nak viszont volt annyi esze, hogy négy évvel később jöjjön a világra, és megkímélje magát ettől a gyötrelemtől.

Na most, a baseball nem a világ legfontosabb témája, legalábbis a PhD-konzulenseim folyton ezt mondogatták. De az alkalmazott módszer a segítségünkre lehet, ha ehhez hasonló kérdésekkel küzdünk – például azt próbáljuk megfejteni, hogy miként alakul ki az emberek pártpreferenciája, szexuális hajlamai, zenei ízlése vagy éppen pénzügyi szokásai. (Különösen az érdekelne, hogy miből fakadnak az öcsém buggyant elképzelései az utóbbi két területen.) A jóslatom az, hogy a felnőttkori viselkedésformáink, érdeklődési köreink jó része, még azok is, amelyekről úgy tartjuk, alapvetően meghatározzák, hogy kik is vagyunk, megmagyarázható olyan esetleges körülményekkel, mint hogy mikor születünk, és mi történt ifjúkorunk bizonyos kulcsfontosságú éveiben.

A politikai preferenciák eredetét már kutatták: Yair Ghitza, a Catalist nevű adatelemző cég vezető kutatója azt a szállóigét tesztelte Andrew Gelmannel, a Columbia Egyetem politológusával és statisztikusával, hogy a több-

ség liberálisként kezdi, majd a korral együtt egyre konzervatívabb lesz. Ezt állítja a híres idézet is, amelyet sokszor Winston Churchillnek tulajdonítanak: „Aki harminc alatt nem liberális, annak nincs szíve, aki pedig harminc felett nem konzervatív, annak nincs esze.”

Ghitza és Gelman hatvan év közvélemény-kutatási adatait böngészte át. Több mint 300 ezer, szavazási preferenciákra vonatkozó adat alapján arra jutottak, hogy a tinédzserek olykor liberális, olykor pedig konzervatív irányba „hajlanak el”. Csakúgy, mint a középkorúak és az idősek.

A két kutató fényt derített arra, hogy a politikai nézetek valójában a baseballdrukkerséghez hasonló módon alakulnak ki. Van egy kulcsfontosságú életszakasz, amely egy életre otthagyja a lenyomatát az emberen. A 14–24 éves korosztályból sokan az éppen aktuális elnök népszerűségére alapozzák nézeteiket. Egy népszerű republikánus vagy egy népszerűtlen demokrata elnök sok fiatal felnőttből republikánust csinál. A népszerűtlen republikánus vagy népszerűtlen demokrata elnökök pedig a demokrata lövészárkba lökhetik az arra fogékony korosztályt.

Az átlagos szavazót általában a síríg elkísérik a kulcsfontosságú évek során kialakult nézetek.

Nézzük meg jobban, hogyan is működik ez: hasonlítsuk össze az 1941-ben világra jött amerikaiakat az egy évtizeddel később születettekkel!

Az első csoportba tartozók Dwight D. Eisenhower, a népszerű republikánus elnök hivatali ideje alatt váltak nagykorúvá. Így ez a generáció erősen a Republikánus Párt felé húzott a hatvanas évek elején, annak ellenére, hogy ekkoriban még harminc alatt voltak. Amikor pedig idősebbek lettek, továbbra is határozottan a republikánus irányba „dőltek”.

A tíz évvel később született amerikaiak – baby boomek – viszont John F. Kennedy, a rendkívül népszerű demokrata elnök, a szintén demokrata, és eleinte szintén népszerű Lyndon B. Johnson, valamint a dicstelenül lemondott, republikánus Richard M. Nixon alatt váltak nagykorúvá. E nemzedék tagjai egész életükben liberális irányba „dőltek”.

Ghitza és Gelman ezen adatokból pontosan meg tudta határozni, hogy mikor alakulnak ki a legnagyobb valószínűséggel a politikai nézetek: tizenennyolc éves korban.

És úgy találták, hogy az efféle bevésozás hatása alapvető jelentőségű. A modelljükön alapuló számítás szerint az „Eisenhower-élmény” mintegy tíz százalékkal növelte az 1941-ben született amerikaiak körében az életfogytig tartó republikánus beállítottságot. A „Kennedy–Johnson–Nixon élmény” pedig hét százalékpontos előnyt jelent a demokratáknak az 1952-es születésűek körében.

Azt már tisztáztam, hogy szkeptikusan viszonyulok a felmérések adataihoz, de itt lenyűgöz a vizsgált válaszok nagy száma. Igaz, ezt a tanulmányt egyetlen, kis mintával dolgozó közvélemény-kutatás alapján nem is lehetett volna elkészíteni. A kutatóknak szükségük volt a sok-sok felmérés adataiból összeállt több százezer megfigyelésre, hiszen csak így láthatták, miként változik a kor előrehaladtával a pártpreferencia.

Az én baseballtanulmányom esetében is kulcsfontosságú volt az adathalmaz mérete. Nem csupán az egyes csapatok szurkolóira kellett ráközelítenem, hanem az egyes korcsoportokra is. Ehhez több millió megfigyelésre van szükség, márpedig a Facebook és más digitális adatforrások rutinszerűen szállítják ezt a mennyiséget.

Itt lépnek működésbe igazán a Big Data méretei. Ahogy egy fotónak rengeteg pixelből kell állnia ahhoz, hogy

tisztán rá tudjunk zoomolni egy apró szeletére, úgy az adathalmaznak is sok-sok megfigyelést kell tartalmaznia ahhoz, hogy ráközelíthessünk az adatok kisebb alcsoportjára – például arra, hogy mennyire népszerű a Mets az 1978-ban született férfiak körében. Egy kis elemszámú, néhány ezres populációval dolgozó felmérés mintája ehhez nem lenne elég nagy.

Ez a Big Data harmadik képessége: lehetővé teszi, hogy ráközelítsünk egy adathalmaz kis szegmentumaira, és ennek alapján megtudjunk valamit magunkról. Nemcsak az életkorra, hanem más dimenziókra is rázoomolhatunk. Ha van elegendő adatunk, megnézhetjük például, hogyan viselkednek az emberek meghatározott kis- és nagyvárosokban. Továbbá óráról órára, vagy akár percről percre nyomon követhetjük, mit művelnek embertársaink.

Ez a fejezet tehát premier plánba hozza az emberi viselkedést.

MI FOLYIK VALÓJÁBAN MEGYÉINKBEN ÉS A VÁROSAINKBAN?

Így, utólag már meglepően hangzik, de tény, hogy amikor Raj Chetty kis kutatócsoportja először jutott hozzá egy elég nagy, mondható adathalmazhoz (az amerikai adónyilvántartáshoz 1996-ig visszamenőleg), nem igazán tudták, ki tudnak-e hozni belőle bármit is. Az amerikai adóhivatal, az Internal Revenue Service (IRS) azért adta át az adatokat a Harvard professzorának, mert abban bíztak, hogy munkatársaival talán fényt tud deríteni az adópolitika hatásaira.

Próbálkozásaik eleinte valóban zsákutcába torkolltak. Bár Big Datával vizsgálták az állami, illetve szövetségi adópolitika következményeit, többnyire ugyanarra az eredményre

jutottak, mint a hagyományos felmérések. A több százmillió IRS-adatpont csak annyit segített, hogy ők talán valamivel precízebbek voltak. De a társadalomtudományban nem minősül komoly teljesítménynek, ha valaki ugyanazt szűri le, mint mások, csak egy kicsit precízebben adatol. A vezető szakfolyóiratok nem ilyen cikkekre utaznak.

Meg aztán meglehetősen időigényes munka volt a tömérdek IRS-adat rendezése és elemzése: az adattengerben fuldokló kutatócsoport lassabban jutott el ugyanazokig a válaszkorig, mint a többiek.

Kezdett úgy festeni a dolog, hogy a Big Data-szkeptikusoknak igazuk van. Vagyis nincs szükség több százmillió ember adataira ahhoz, hogy megértsük az adópolitikát: bőven elég egy tízezer fős felmérés. Chettynek és csapatainak érthető módon kedvét szegte a dolog.

De aztán végre-valahára rádöbrentek, hogy mi volt a hiba. „A Big Data nem arról szól, hogy ugyanazt csináljuk, mint a felméréseknél, csak több adattal” – magyarázza Chetty. Kis adatmennyiségre szabott kérdéseket tettek fel az irdatlan adattömegnek. „A Big Datának lehetővé kell tennie, hogy egészen más modelleket használjunk, mint egy felmérésnél” – teszi hozzá Chetty. „Ráközelíthetünk például kisebb földrajzi egységekre.”

Tehát a százmillió adathalmaz segítségével meg tudták figyelni a kisebb-nagyobb városokban, városnegyedekben, környékeken kirajzolódó mintázatokat.

Harvardi doktoranduszként ott ültem a szeminárium szobájában, ahol Raj Chetty professzor bemutatta az adónyilvántartás alapján végzett kutatás kezdeti eredményeit. A társadalomtudósok műveiben gyakran esik szó megfigyelésekről, vagyis arról, hogy hány adatpont áll rendelkezésükre. Ha például valaki egy 800 fős felmérés eredményei alapján dolgozik, akkor azt mondja, hogy „nyolcszáz

megfigyelésünk van”. Ha pedig laboratóriumi kísérletet végez 70 emberrel, akkor értelemszerűen azt mondja, hogy „hetven megfigyelésünk van”.

„Egy egész két tized milliárd megfigyelésünk van” – jelentette be akkor pókerarccal a professzor. A hallgatóság idegesen kacarászott.

Raj Chetty és szerzőtársai ott, a szemináriumi szobában, majd később a cikksorozatukban fontos felismerésekkel gazdagítottak minket arról, hogyan is működik Amerika.

Vegyük például a következő kérdést: Amerika tényleg a korlátlan lehetőségek hazája? Más szóval: van-e esélye valakinek arra, hogy meggazdagodjon, ha a szülei nem gazdagok?

A kérdés megválaszolásának hagyományos módja az, hogy megnézünk egy reprezentatív amerikai mintát, és összevetjük más országok hasonló adataival.

Íme, néhány ország esélyegyenlőségi adata. A kérdés a következő volt: mekkora eséllyel juthat el valaki a jövedelemelosztás szerinti alsó 20 százalékból a felső 20 százalékba?

ANNAK ESÉLYE, HOGY A SZEGÉNY SZÜLŐK GYERMEKE
MEGGAZDAGODJON (VÁLOGATOTT ORSZÁGOK ADATAI)

Egyesült Államok	7,5
Egyesült Királyság	9,0
Dánia	11,7
Kanada	13,5

Mint láthatjuk, Amerika *nem* teljesít valami jól.

Csakhogynem ebben az egyszerű elemzésben elvessz a lényeg. Amikor azonban Chetty csapata „rázoomolt” a kisebb földrajzi egységekre, kiderült: az esélyek erősen függenek attól, hogy az illető az Egyesült Államok mely részén született.

ANNAK ESÉLYE, HOGY A SZEGÉNY SZÜLŐK
GYERMEKE MEGGAZDAGODJON
(AZ EGYESÜLT ÁLLAMOK KÜLÖNBÖZŐ RÉSZEINEK ADATAI)

San José, Kalifornia	12,9
Washington DC	10,5
Országos átlag	7,5
Chicago, Illinois	6,5
Charlotte, Észak-Karolina	4,4

Az Egyesült Államok egyes részein ugyanakkora eséllyel lehet sikeres egy szegény gyerek, mint a világ más fejlett országaiban. Az Egyesült Államok más részein viszont kisebb eséllyel lehet sikeres egy szegény gyerek, mint a világ más fejlett országaiban.

Egy kis elemszámú felmérésben, amelyben csak néhány charlotte-i és San José-i lakos vesz részt, efféle mintázatok nem rajzolódhatnak ki, így nem is tudnánk a fenti módon ráközelíteni semmire.

Chetty csapata azonban még ennél tovább is zoomolhatott. Amerika minden lakójának adatai ott voltak a kezükben. Még azok kisebb csoportjait is vizsgálni tudták, egyik városból a másikba költöztek. Megnézhették, milyen hatással van az emberek kilátásaira, ha átteszik a székhelyüket New Yorkból Los Angelesbe, Milwaukee-ból Atlantába vagy éppen San Joséból Charlotte-ba. Ez pedig lehetővé tette, hogy ne csak a korrelációt, hanem az oksági összefüggést is vizsgálják. (A kettő közti különbségről a következő fejezetben ejtek szót.) Kiderült, hogy ha az ember a fejlődése meghatározó éveiben a megfelelő városba költözik, az valóban jelentősen javítja az esélyeit.

A korlátlan lehetőségek hazája-e tehát Amerika?

A válasz: sem nem igen, sem pedig nem. A válasz: bizonyos országrészekben igen, máshol pedig nem.

A szerzők szavaival élve: „Az Egyesült Államokat sokkal inkább különböző társadalmak együtteseként írhatnánk le. Ezek egy része valóban a lehetőségek hazája, ahol jelentős a nemzedékek közötti mobilitás, másutt viszont csak kevés gyerek tud kitörni a szegénységből.”

Mi a helyzet tehát az Egyesült Államok azon részein, ahol magas a jövedelmi mobilitás? Mitől van az, hogy bizonyos helyeken kevésbé „lejt a pálya”, vagyis egy szegény gyereknek is lehet egész jó élete? Azokon a területeken, ahol többet költenek az oktatásra, jobb esélyek nyílnak a szegény gyerekek előtt. Jobban teljesítenek az olyan területek is, ahol több vallásos ember él, és kisebb a bűnözés. Rosszabbul teljesítenek az olyan területek, ahol több fekete él. Ez érdekes módon nem is csak a fekete gyerekek esélyeire hat ki, hanem az ott élő fehér gyerekeikéire is. Ugyanez a helyzet azokkal a területekkel is, ahol sok az egyedülálló anya. Ez pedig szintén kihat e területek házasságban élő lakóinak gyerekeire is. Az eredmények részben azt sugallják, hogy számít a kortárs csoport is. Ha egy gyerek barátainak rossz a hátterük és kevés a lehetőségük, akkor ő is nehezebben tud kitörni a szegénységből.

Az adatok tanúsága szerint Amerika egyes részei több esélyt nyújtanak a gyerekeknek arra, hogy maguk mögött hagyják a szegénységet. De vajon hol a legjobb az emberek esélye arra, hogy megszökjenek „a Nagy Kaszás” elől?

Szívesen gondolunk úgy a halálra, mint az egyenlőség bajnokára. A halált senki sem kerülheti el: sem a koldus, sem a királyfi, sem a hajléktalan, sem Mark Zuckerberg. Mindenki meghal.

Viszont a gazdagok, ha el nem kerülhetik is a sorsukat,

az adatok tanúsága szerint manapság már egész jól el tudják napolni a dolgot. Azok az amerikai nők például, akik a jövedelemeloszlás szerinti legfelső egy százalékba tartoznak, tíz évvel tovább élnek, mint a legalsó egy százalékba tartozók. A férfiaknál pedig tizenöt év a különbség.

Vajon miként változnak ezek a mintázatok az Egyesült Államok különböző területein? Befolyásolja az ember várható élettartamát az, hogy hol él? És vajon ugyanakkora a földrajzi alapú különbség a gazdagok, mint a szegények esetében? Raj Chetty csapata a kisebb földrajzi egységekre ráközelítve ezekre a kérdésekre is megadta a választ.

Érdekes módon a leggazdagabb amerikaiak várható élettartamára alig-alig hat ki a lakóhelyük. Akit felvet a pénz, az számíthat rá, hogy hozzávetőlegesen 89 évig fog élni, ha nő, és 87 évig, ha férfi. A gazdagok mindenütt egészségesebb szokásokat vesznek fel: átlagosan többet mozognak, jobban táplálkoznak, kevesebbet dohányoznak, és sokkal kevésbé fenyegeti őket az elhízás. A gazdagoknak telik futópadra, organikus avokádóra, jogatanfolyamra. És mindezeket az ország bármelyik részén megvásárolhatják.

A szegények esetében viszont más a történet. A legszegényebb amerikaiak várható élettartama óriási mértékben függ a lakóhelyüktől. Nem kevesebb mint öt évvel hosszabbíthatja meg az életüket, ha a megfelelő helyen élnek.

És vajon miért van az, hogy bizonyos helyek elszegényedett lakosainak ennyivel több jut? Milyen tulajdonságokkal rendelkeznek azok a városok, amelyekben a szegények a legtovább élnek?

A városok alábbi négy jellemzője közül három nem korrelál a várható élettartammal, egy viszont igen. Lássuk, kitalálja-e a kedves Olvasó, hogy melyik számít!

MITŐL ÉLNEK JÓVAL TOVÁBB EGY VÁROS SZEGÉNYEI?

A városban magas a vallásosság szintje.
A városban alacsony a légszennyezettség szintje.
A városban magasabb az egészségbiztosítással rendelkezők aránya.
Sok gazdag ember él a városban.

Az első három tényező – a vallásosság, a környezetszennyezés, valamint az egészségbiztosítási lefedettség – nem korrelál a szegények hosszabb élettartamával. Chetty és társai szerint tehát a negyedik változó a megoldás: a városban élő gazdagok száma. Azaz minél több valahol a gazdag ember, annál tovább élnek az ottani szegények. A New York-i szegények például jóval tovább élnek, mint a detroitiak.

De vajon miért jelzi előre ilyen jó hatásokkal a szegények várható élettartamát a gazdagok jelenléte? David Cutlernek, a tanulmány egyik szerzőjének, aki egyúttal az egyik konzulensem, van egy hipotézise, amely egyelőre csupán spekulatív alapokon nyugszik. A jelenség mögött részben talán az állhat, hogy az életmód ragályos.

Számos kutatás igazolja, hogy az életmód ragályos. Ilyen alapon elképzelhető, hogy a szegények átvesznek bizonyos szokásokat a közelükben élő gazdagoktól. E szokások egy része, például a fellengzős szóhasználat, aligha befolyásolja az egészségüket, más szokások viszont, például a testedzés, egyértelműen pozitív hatásúak. Ami azt illeti, a gazdagok közelében élő szegények valóban többet mozognak, kevesebbet dohányoznak, és kisebb valószínűséggel híznak el.

Az én személyes kedvencem Raj Chetty csapatának az a tanulmánya, amelyben az IRS óriási adathalmazát felhasználva azt vizsgálták, miért csálnak egyesek adót, mások pedig miért nem. E tanulmány ismertetése valamivel komplikáltabb feladat.

A lényeg a következő: ha valaki önfoglalkoztatónak minősül, és egy gyermeke van, egy pofonegyszerű trükkkel maximalizálhatja az államtól kapott pénzét. Ha egy adott évben pontosan 9000 dollár adóköteles jövedelmet vall be, akkor az állam kiállít neki egy csekket 1377 dollárról – ez az úgynevezett Earned Income Tax Credit (EITC), azaz a megkeresett jövedelem utáni adójóváírás. Ez a dolgozó szegényeknek járó juttatás, amellyel az állam kiegészíti a munkavállaló összes, a járulékok levonása utáni jövedelmét. Ha valaki 9000 dollárnál többet vall be, akkor több járulékot fog fizetni; ha viszont kevesebbet, akkor az EITC összege csökken. Így hát évi 9000 dollár adóköteles jövedelem a tuti.

És történetesen – hát nem hihetetlen? – a 9000 dollár a leggyakoribb összeg, amit az egygyerekes önfoglalkoztatók bevallanak!

Direkt úgy alakította volna ez a sok ember a munkarendjét, hogy év végén pont ez az ideális összeg jöjjön ki? Dehogy. Amikor szűrőpróbaszerűen ellenőrizték őket – ami nagyon ritkán fordul elő –, szinte mindig az derült ki, hogy a tényleges jövedelmük a közelében sem járt a 9000 dollárnak: vagy jóval többet kerestek, vagy jóval kevesebbet.

Vagyis adót csaltak: azt az összeget írták a bevallásukba, amely után az állam a legkövetkező csekket fizeti.

Felmerül a kérdés, hogy mennyire volt jellemző ez a fajta adócsalás, illetve hogy az érintett csoporton belül kik követik el a legnagyobb valószínűséggel.

Chetty és kollégái kiderítették, hogy óriási különbségek vannak az országban belül. Miami-ban például döbbenetesen sok egygyerekes önfoglalkoztató vallott be pontosan 9000 dollárt: az e kategóriába tartozók nem kevesebb mint 30 százaléka. Philadelphiában mindössze két százalékuk tett így.

Vajon mi jelzi előre, hogy ki fog csalni? Milyen sajátosság van azokban a városokban, térségekben, ahol a leg-

magasabb, illetve a legalacsonyabb a vétkezők száma? Ha korreláljuk az ilyen esetek számarányát más, városi szintű szociológiai adatokkal, kiderül, hogy a gyakoriságot két tényező befolyásolja igen erősen: az, hogy az adott területen magas az EITC-re jogosultak, illetve az adószakértők koncentrációja.

Hogy mit jelez az említett két tényező? Chetty és szerzőtársai erre is adtak magyarázatot. Az efféle adócsalás fő mozgatórugója az információ volt.

Az egygyerekes önfoglalkoztatók többsége egyszerűen nem tudta, hogy a 9000 a bűvös szám. De ha valaki olyanok közelében él, akik tudhatnak erről – vagyis a szomszédai is hasonló cipőben járnak, vagy sok a környéken az adótanácsadó –, akkor ugrásszerűen megnő az esélye arra, hogy a fülébe jusson a hír.

Chetty és kollégái további bizonyítékokat is találtak arra, hogy az EITC-vel való ügyeskedés fő hajtóereje a tudás. Azok, akik olyan helyről, ahol kevesen csinálnak ilyet, átköltöztek egy EITC-csalás sújtotta övezetbe, eltanulják és alkalmazzák a 9000 dolláros trükköt. Így az adójóváírással való trükközés idővel régióról régióra terjedt. Az adócsalás tehát fertőz, mint valami vírus.

Most pedig álljunk meg egy percre, és gondoljuk át, mire is derített fényt ez a tanulmány? Azt demonstrálta, hogy ha meg akarjuk tudni, ki fog adót csalni, nem azt kell megállapítanunk, hogy ki becsületes és ki nem. Hanem azt, hogy ki tudja, hogyan kell adót csalni, és ki nem.

Ha tehát valaki azt mondja önnek, hogy ő soha, de soha nem csalna adót, akkor jó eséllyel – igen, eltalálta – hazudik. Raj Chetty kutatási eredményei szerint sokan megtennék, ha tudnák, hogyan kell.

Ha tehát valaki adót akar csalni (amit természetesen nem javaslok!), akkor érdemes adószakértők vagy adó-

csalók között élnie, akik ki tudják okosítani. Aki pedig azt akarja, hogy a gyerekei világhírűek legyenek, az hol lakjon? Mivel ma már roppant aprólékosan, egy-egy kisebb adathalmazra ráközelítve tudunk dolgozni, talán ezt a kérdést is sikerül megválaszolnunk.

Érdekelt, hogy vajon honnan származnak a legsikeresebb amerikaiak, ezért egy szép napon elhatároztam, hogy le-töltöttem a Wikipédiát. (Ma már ezt is lehet.)

Némi kódolgatás után a birtokomban voltak több mint 150 ezer olyan ember adatai, akiket a Wikipédia érdekesnek talált egy külön szócikkre. Az adathalmazban benne volt a lexikonban szereplő személyek születési helye, születési dátuma, foglalkozása és neme. Mindezt egybevettem az Országos Egészségügyi Statisztikai Központ megyénkenti születési adataival. Így az ország minden megyéjére ki tudtam számolni, hogy mekkora eséllyel kerül be a Wikipédiába, aki ott született.

De vajon megbízható mércéje-e a kiugró teljesítménynek az, hogy valakinek van-e Wikipédia-oldala? Nyilvánvaló, hogy csak megszorításokkal. A szerkesztőgárda jórészt fiatal férfiakból áll, amiktől torzulhat a minta. Ráadásul az ismeretségnek akadnak olyan okai is, amelyek valójában nem különösen értékesek. Ted Bundy például azzal érdemelte ki a Wikipédia-szócikket, hogy fiatal nők tucatjait gyilkolta meg. De a bűnözőket végül sikerült kigyomlálnom, így a hiányuk nem befolyásolta különösebben az eredményt.

Az 1946 és 1964 között született baby boomerekre korlátoztam a mintát, mert nekik már bőven volt idejük arra, hogy híressé váljanak. Ennek a nemzedéknek minden 2058-adik tagját érdemesítették Wikipédia-szócikkre. A lexikonban szereplők mintegy 30 százaléka a művészek terén vagy a szórakoztatóiparban tűnt ki, 29 százalé-

kuk a sportban, 9 százaléka a politikában, 3 százaléka pedig a tudományban.

Először az szűrt szemet, hogy a kiugró siker valószínűsége földrajzilag rendkívül változatos képet mutat, legalábbis a Wikipédia szerint. A hírnévre mutató esély erősen függ attól, hogy hol született az ember.

A kaliforniai születésű baby boomerek közül minden 1209-edik került fel a Wikipédiára, a nyugat-virginiai születésűek közül viszont csak minden 4496-odik. Ha pedig rázoomolunk a megyénkénti adatokra, még árulkodóbb az eredmény. A massachusettsi Suffolk megyéből, ahol Boston is található, minden 748-adik baby boomer szerepel a Wikipédián, de vannak megyék, ahol a sikerráta ennek csupán a huszada.

Vajon mitől ennyivel jobb az ország egyes részei, miért termelik futószalagon Amerika „motorjait”? Miután átvizsgáltam az élváros megyéket, kiderült, hogy mind-egyikre ráillik az alábbi két paraméter valamelyike.

Először is – amin meglepődtem –, sok ilyen megyében van jelentős egyetemváros. Ha számomra ismeretlen név bukkant fel az élbolyban – például a michigani Washtenaw –, szinte mindig kiderült, hogy azt a megyét egy klaszszikus egyetemváros dominálja – Washtenaw megye esetében például Ann Arbor. Az alábbi egyetemvárosok által fémjelzett megyék mind bekerültek a legfelső 3 százalékba: Madison (Wisconsin állam), Athens (Georgia), Columbia (Missouri), Berkeley (Kalifornia), Chapel Hill (Észak-Karolina), Gainesville (Florida), Lexington (Kentucky) és Ithaca (New York).

Vajon miért van ez így? Felettébb valószínű, hogy részben a genetika miatt, hiszen a professzorok, doktoranduszok fiai-lányai jellemzően okosak (Ímí roppant hasznos

tulajdonság, ha a játszma tétje a kiugró siker). És valóban: ha egy adott földrajzi területen sok a diplomás, az előre-títi az ott született emberek sikereit.

De nyilvánvaló, hogy valami más is meghúzódik az egyetemvárosok sikere mögött: az újító szellemnek való korai kitettség. Az egyik olyan terület, amelyen az egyetemvárosok a leginkább sikeresek, a zene. Aki ilyen helyen nő fel, azt sok különleges, egyedi hatás éri: exkluzív koncertekre jár, nem kommersz rádióadókat hallgat, sőt még független lemezboltot is láthat közelről. És mindez nem csak a művészetekre érvényes: az egyetemvárosok a neves üzletemberek keltetője is. Talán nekik is segítségükre van a legfrissebb szellemű alkotásoknak, eszméknek való kitettség.

Az egyetemvárosok sikere nemcsak országgrészeken ível át, hanem a faji határokon is. Az afroamerikaiak a sportot leszámítva feltűnően alulreprezentáltak a Wikipédián, különösen az üzleti szférában és a természettudományban. Kétségtelen, hogy ebben komoly szerepet játszik a diszkrimináció is. De van egy kis megye, ahol az 1950-ben születettek 84 százaléka fekete, és ez a megye mégis közel ugyanannyi hírességet adott a világnak, mint a rangsor élén állók.

Az alabamai Macon megye alig 13 ezer baby boomere közül tizenöten kerültek fel a Wikipédiára, vagyis minden 852-edik. Kivétel nélkül feketék. A tizenötből tizennégyen Tuskegee szülöttei: ez a város a Booker T. Washington által feketék számára létrehozott Tuskegee Egyetem otthona. Bírák, írók, természettudósok szerepelnek a listán. Egy Tuskegee-ben született fekete gyerek ugyanakkora eséllyel lett híres valamiben, ami nem sport, mint a fehér gyerek, aki fehér többségű elit egyetem szomszédságában jött világra.

A második paraméter, amely alapján valószínűsíthető

egy megye szülötteinek sikere, az, hogy az adott megyében van nagyváros.

San Francisco megye, Los Angeles megye és New York City egyaránt kiváló esélyt kínál a Wikipédia-szócikkre. (New York City öt megyéjét összevontam, mivel a Wikipédia-szócikkekben sokszor nem tüntetik fel, hogy ki melyik kerületben született.)

A nagyvárosi környezet rendszerint bőséggel kínál mintákat a sikeres életre. Hogy lássuk, mekkora érték, ha az ember egy mesterség sikeres művelőinek közelében lehet, hasonlítsuk össze New York Cityt, Bostont és Los Angeles-t. A három közül New York City produkálja arányosan a legtöbb jelentős újságíró, Boston a legtöbb jelentős természettudóst, Los Angeles pedig a legtöbb jelentős színészt. (Ne feledjük: azokról beszélünk, akik ott születtek, nem azokról, akik odaköltöztek!) És mindez akkor is igaz, ha nem számítjuk azokat, akiknek már a szülei is hírnévre tettek szert az adott területen.

A kertvárosias megyék, ha nincs a területükön jelentős egyetemváros, sokkal rosszabbul teljesítenek, mint a nagyvárosias megyék. Az én szüleim, ahogy a baby boomer nemzedékből oly sokan, kiköltöztek a zsúfolt járdák világából a csendes, árnyas utcák világába – konkrétan Manhattanból a New Jersey-i Bergen megyébe –, hogy ott neveljék fel három gyermeküket. Elképzelhető, hogy ez hiba volt, legalábbis abból a szempontból, hogy mekkora eséllyel lesznek híresek a gyerekeik. A New York Cityben született gyerekeknek 80 százalékkal nagyobb az esélyük arra, hogy felkerüljenek a Wikipédiára, mint a Bergen megyében születetteknek. Ez persze csupán korreláció, de mindenképpen azt sugallja, hogy ha valaki nagy gondolatok szomszédságában nő fel, az többet használ neki, mint ha egy szép nagy udvaron szaladgálna kiskorában.

A gyerekkori környezet hatásai talán még határozottabban kirajzolódnának, ha több adatom lenne arról, ki hol élt gyerekkorában, hiszen sokan nem abban a megyében nőnek fel, ahol megszülettek.

Az egyetemvárosok és a metropoliszok kedvező hatása eléggé szembeötlő. Én azonban egy kicsit jobban belemélyedtem a dologba, és elvégeztem egy valamivel szofisztikáltabb empirikus elemzést is.

Amely kimutatta, hogy van még egy lényeges változó a Wikipédián való megjelenés vonatkozásában: a bevándorlók számaránya. Minél több külföldi születésű lakosa van a megyének, ahol az ember született, annál több esélye van arra, hogy később sokra vigye. (Ezt kapd ki, Donald Trump!) Ha van két megye, ahol hasonló a nagyvárosi, illetve egyetemvárosi lakosság számaránya, akkor a több bevándorlót felvonultató megye termel ki több jeles személyiséget. Erre vajon mi a magyarázat?

Úgy tűnik, hogy a dolog jelentős részben közvetlenül a bevándorlók gyerekeinek tulajdonítható. Alapos kutatást végeztem a száz leghíresebb fehér baby boomereknek a Massachusetts Institute of Technology (MIT) Pantheon-projektjében rögzített életrajza kapcsán, amely szintén Wikipédia-adatokkal dolgozik. Egyrészt az derült ki, hogy a többség vállalkozó volt, másrészt pedig az, hogy legalább tizenhármuknak (köztük Oliver Stone-nak, Sandra Bullocknak és Julianne Moore-nak) külföldön született az édesanyja. Ez a 13 százalékos arány több mint háromszorosa az ugyanezen időszakban mért országos átlagnak. (A százas toplistán szereplők közül sokaknak az apja volt bevándorló, így Steve Jobsnak és John Belushinak is, de ezt az adatot nehezebb összevetni az országos átlaggal, mert a születési bizonyítvány nem mindig tartalmaz információt az apáról.)

És vajon mely változók *nincsenek* hatással a későbbi kiemelkedő sikerekre? Az egyik meglepett, nem is kicsit: kiderült, hogy ebből a szempontból nem számít, mennyit költ egy állam az oktatásra. Ha két államban hasonló a nagyvárosi környezetben élő népesség számaránya, akkor az oktatási célú kiadások nagysága nem korrelál a kibocsátott jelentős írók, művészek, vezető üzletemberek számarányával.

Érdemes összevetni az én Wikipédia-tanulmányomat Raj Chetty csapatának egyik tanulmányával, amelyről már esett szó. Chetty és társai, ugyebár, azt próbálták kideríteni, hogy hol könnyebb „felkapaszkodni” a felső középosztályba, én pedig arra kerestem a választ, hogy hol könnyebb hírnevet szerezni. Egészen megdöbbentő, hogy mennyire eltérő eredményre jutottunk.

Mint kiderült, ha valahol sokat költenek az oktatásra, az elősegíti, hogy egy gyerek felküzdje magát a felső középosztályba. Azt viszont vajmi kevésbé segíti elő, hogy egy gyerekből jelentős író, művész vagy üzletember legyen. A kiugróan sikeres emberek közül rengetegen utáltak az iskolát. Egyesek ki is hullottak a rendszerből.

Chetty és csapata arra jutott, hogy New York City nem különösebben jó választás, ha azt akarjuk, hogy gyermekünk később a felső középosztály tagja legyen. Én pedig arra jutottam, hogy remek választás, ha azt akarjuk, hogy esélye legyen a hírnévre.

Ha megnézzük a sikert előmozdító tényezőket, kezd értelmet nyerni a megyék közti hatalmas különbség. Számos megyében kombináltan vannak jelen a siker fő kellei. Térjünk vissza ismét Bostonhoz! A városban számos egyetem működik, így szinte forr a levegő az innovatív ötletektől. Továbbá nagyvárosról beszélünk, ahol sok-sok ki-

váló elme mutat példát az ifúságnak. Egy ilyen hely pedig vonzza a bevándorlókat, akiknek a gyerekei roppant motíváltan igyekeznek majd hasznosítani a tanultakat.

De mi van akkor, ha egy megye nem rendelkezik ilyen adottságokkal? Mindenképpen arra van kárhoztatva, hogy kevesebb szupersztárt adjon a világnak? Nem feltétlenül. Van egy másik lehetséges út is: a szélsőséges specializáció. Jó példa erre a minnesotai Roseau megye. Ebben a kis területű, vidékies megyében kevés külföldi él, ráadásul jelentős egyeteme sincs, mégis minden 740-edik ott született ember felkerült a Wikipédiára. Hogy mi Roseau megye titka? Az, hogy mind a kilenc híressége jégkorongozó, amiben nyilván közrejátszottak a megyében működő, világszínvonalú ifúsági és középiskolai hokiprogramok.

Az lenne tehát a lényeg – már ha valaki nem pont hokisztárt akar faragni a gyerekekből –, hogy költözzünk Bostonba vagy Tuskegee-be, mert leendő utódaink szempontjából ez a legelőnyösebb? Az biztos, hogy baj nem lehet belőle. De vannak itt általánosabb érvényű tanulságok is.

A közgazdászok és a szociológusok rendszerint arra összpontosítják a figyelmüket, hogy miként kerülhetnénk el a bajokat, úgymint a szegénységet és az elharapózó bünnözést. Egy jó, sőt kiváló társadalom azonban nemcsak azt a célt tűzi maga elé, hogy minél kevesebb ember szakadjon le, hanem azt is, hogy minél több embernek segítsen kiemelkedni. Azzal, hogy „rázoomoltunk” a földrajzi helyekre, ahol a híres amerikaiak százazrei látták meg a napvilágot, talán sikerült megvetnünk egy jövőbeni nemzetstratégia alapjait. A stratégia részét képezhetné például a bevándorlás bátorítása, valamint az egyetemek és a művészetek támogatása.

Általában az Egyesült Államokat tanulmányozom. Ezért amikor földrajzi alapú zoomolásban gondolkodom, kisebb-nagyobb városaink, megyéink lebegnek a szemem előtt: olyan helyek, mint az alabamai Macon megye, vagy éppen a minnesotai Roseau megye. De az internetes adatoknak van egy másik hatalmas (és egyre növekvő) előnyük is: az, hogy könnyen gyűjthetünk információkat a világ minden tájáról, és megnézhetjük, miben térnek el egymástól az egyes országok. Az adattudósok előtt pedig megnyílik a lehetőség, hogy lábujjhegyen beosonhassanak az antropológia vadászterületére.

Íme, egy kissé esetleges téma, amellyel mostanában foglalkoztam: hogyan játszódik le a világ különböző tájain a terhesség? Megvizsgáltam a várandós nők Google-kereséseit, és elsőként az tűnt fel, hogy mindenütt döbbenetesen hasonló fizikai tünetekre panaszkodnak.

Leteszteltem, hogy milyen gyakorisággal kerestek rá a különböző tünetekre, a „terhes” kifejezéssel párosítva. Példának okáért hányszor társítják a „terhes” szóhoz a „hányinger”, a „hátfájás” vagy a „székrekedés” kifejezéseket. Kanada tünetei nagyon közel állnak az amerikaiakhoz. És többé-kevésbé hasonló tünetekre kerestek rá például Nagy-Britanniában, Ausztráliában és Indiában is.

A terhes nők világszerte láthatóan ugyanarra áhítoznak. Az Egyesült Államokban a leggyakoribb, a témához kapcsolódó keresőkifejezés az, hogy „fagyira vágyakozni terhesség alatt”. A következő négy leggyakoribb a só, az édességek, a gyümölcs és a csípős étel. Ezek a vágyak Ausztráliában sem különböznek nagymértékben: ott só, édesség, csoki, fagy, gyümölcs a sorrend. És mi a helyzet Indiával? Ott is hasonló a történet: csípős kaja, édesség, csoki, só, fagy. Elmondhatjuk, hogy az ötös toplista az összes vizsgált országban meglehetősen hasonló.

Az előzetes adatok tehát arra utalnak, hogy a világ egyetlen pontján sem bukkant még fel olyan étrend vagy környezeti hatás, amely drasztikusan megváltoztatná a terhesség fizikai élményét.

A terhesség alatt megjelenő gondolatok viszont határozottan különböznek.

Kezdjük az azzal kapcsolatos kérdésekkel, hogy mit tehet, illetve nem tehet meg biztonsággal egy várandós nő! Az Egyesült Államokban a következő keresések a leggyakoribbak: szabad-e egy terhes nőnek „garnélarákot enni”, „bort inni”, „kávét inni”, „Tylenolt szedni”?

Az efféle aggályok tekintetében a többi ország nemigen mutat hasonlóságot sem az Egyesült Államokkal, sem pedig egymással. Hogy szabad-e a terhes nőknek „bort inni”, az sem Kanadában, sem Ausztráliában, sem Nagy-Britanniában nem szerepel a keresések tízes toplistáján. Ausztráliában főként a tejtermékek, azok közül is elsősorban a krémsajt fogyasztása vált ki aggodalmat, Nigériában pedig, ahol a népesség 30 százaléka használja az internetet, az a leggyakoribb kérdés, hogy ihat-e egy terhes nő hideg vizet.

Hogy jogosak-e ezek az aggályok? Az attól függ. Arra, hogy a terhes nőknél fokozott a pasztörizetlen sajt okozta listeriózis kockázata, bőven van bizonyíték. Kimutatták azt is, hogy a túlzott alkoholfogyasztás negatív következményekkel járhat a gyermek szempontjából. A világ egyes részein úgy tartják, hogy a hideg víztől tüdőgyulladást kaphat a baba, de nem tudok róla, hogy ez orvosilag alátámasztott tény lenne.

Az, hogy a feltett kérdések közt ilyen hatalmas különbségek vannak, alighanem azzal magyarázható, hogy a világ minden táján igen sokféle információ zúdul az emberekre: tudományos teóriák, régi hiedelmek, a szomszédság locsogása. Így a nők nehezen tudják eldönteni, hogy mire

összpontosítsák a figyelmüket, vagyis hogy mire keressenek rá a Google-on.

Egyértelmű különbségeket tapasztalunk akkor is, ha megnézzük a „hogyan kell [...] a terhesség alatt?”-témájú kereséseket. Az Egyesült Államokban, Ausztráliában és Kanadában a leggyakoribb ilyen jellegű kérdés a következő: „hogyan kell megelőzni a terhességi csíkokat a terhesség alatt”? Ghánában, Indiában és Nigériában viszont az ötös toplistán sem szerepel a terhességi csíkok elhárítása: ezekben az országokban inkább a szex és az alvás érdekli a várandós nőket.

AZ ÖT LEGGYAKORIBB „HOGY KELL [...] A TERHESSÉG ALATT?”-TÉMÁJÚ KERESÉS (SORRENDEN)

EGYESÜLT ÁLLAMOK	INDIA	AUSZTRÁLIA	NAGY-BRITANNIA	NIGÉRIA	DÉL-AFRIKA
megelőzni a terhességi csíkokat	aludni	megelőzni a terhességi csíkokat	fogyni	szexelni	szexelni
fogyni	csinálni a szexet	fogyni	megelőzni a terhességi csíkokat	fogyni	fogyni
szexelni	szexelni	elkerülni a terhességi csíkokat	elkerülni a terhességi csíkokat	szeretkezni	megelőzni a terhességi csíkokat
elkerülni a terhességi csíkokat	szex	aludni	aludni	egészségesnek maradni	aludni
fittnak maradni	vigyázni magunkra	szexelni	szexelni	nem háyni	nem háyni

AZ ÖT LEGGYAKORIBB „LEHET-E [...] A TERHESSÉG ALATT?”-TÉMÁJÚ KERESÉS

EGYESÜLT ÁLLAMOK	garnélarákot enni	bort inni	kávét inni	Tylenolt szedni	szusit enni
NAGY-BRITANNIA	rákot enni	füstölt lazacot enni	sajttortát enni	mozzarella-t enni	majonézt enni
AUSZTRÁLIA	krémsajtot enni	rákot enni	szalonnát enni	tejfölt enni	fetasajtot enni
NIGÉRIA	hideg vizet inni	bort inni	kávét inni	szexelni	moringát (ehető növényt) fogyasztani
SZINGAPÚR	zöld teát inni	fagyit enni	duriánt enni	kávét inni	ananászt enni
SPANYOL-ORSZÁG	pástétomot enni	sonkát enni	paracetamolt (fájdalomcsillapító) szedni	tonhalat enni	napozni
NÉMET-ORSZÁG	repülőre ülni	szalámit enni	szau-názni	mézet enni	mozzarella-t enni
BRAZÍLIA	haját festetni	Dipironát (fájdalomcsillapító) szedni	paracetamolt szedni	biciklizni	repülőre ülni

Ennél nyilván sokkal többet is megtudhatunk, ha ráközelítünk a világ különböző szegleteinek egészségügyi, illetve kulturális sajátosságaira. De már az én előzetes kutatásaim is sejteni engedik, hogy mit fog mondani nekünk

a Big Data. Azt, hogy mi, emberek, gyámoltalanabbak vagyunk, mint gondolnánk, ha biológiai korlátaink meghaladásáról van szó. Azt viszont, hogy mivel járnak ezek a korlátok, meglepően sokféleképpen értelmezzük.

HOGYAN TÖLTJÜK KI PERCEINKET, ÓRÁINKAT?

„Egy fiatalember kalandjai, akit főként a nők meggyalázása, az extrém erőszak és Beethoven érdekeli.”

Így hirdették Stanley Kubrick sokat vitatott filmjét, a *Mechanikus narancsot*. Az ifjú, fiktív főhős, Alex DeLarge vérfagyasztó közönnyel követi el sokkolóan erőszakos tetteit. Az egyik leghírhedtebb jelenetben torkaszakadtából bömböli az „Ének az esőben”-t, miközben megerőszakol egy nőt.

Szinte azonnal elkezdtek beérkezni a hasonló incidensekről szóló jelentések. Ami azt illeti, egy csapatra való férfi ugyanezt a dalt énekelve erőszakolt meg egy tizenhét éves lányt. A filmet több európai országban betiltották, az Amerikában vetített verzióból pedig kivágták a legdurvább jelenetek egy részét.

Valóban számos példa van arra, hogy a való élet utánozza a művészetet: ilyenkor az embereket mintha hipnotizálná az, amit a mozivásznnon láttak. A *Színek* című gengszterfilm egyik vetítését például heves lövöldözés követte, a *New Jack City* című gengszterfilm bemutatóját pedig zavargások.

A legfelkavaróbb eset alighanem a *Pénzvonat* című akcióvígjáték bemutatója után négy nappal történt. Néhány férfi öngyújtófolyadékkal lángra lobbantott az aluljáróban egy pénztárfülkét. Szinte hibátlanul utánózták a film egyik

jelenetét, csak az volt a különbség, hogy a filmben megmenekült a pénztáros, a valóságban viszont halálra égett.

Pszichológiai kísérletek is igazolják, hogy az alanyok egy erőszakos film hatására felfokozott indulatokról számolnak be, még akkor is, ha nem utánozzák pontosan valamelyik konkrét jelenetet.

Vagyis mind az anekdotikus történetek, mind a kísérletek azt sugallják, hogy az erőszakos filmek erőszakos viselkedést válthatnak ki. De vajon mekkora a hatásuk valójában? Évtizedenként két gyilkosságról beszélünk, vagy évi több száz gyilkosságról? Erre már sem az anekdoták, sem a kísérletek nem tudnak választ adni.

Két közgazdász, név szerint Gordon Dahl és Stefano DellaVigna úgy döntött, utánanézz, képes-e erre a Big Data. Összegyűrtak három óriási adathalmazt, amelyek az 1995 és 2004 közötti időszakot ölelték fel: az FBI óránkénti bűnözési adatait, a mozijegy-eladási adatokat, valamint a kids-in-mind.com weboldal filmadatbázisát, amely tízes skálán pontozza az erőszak mértékét.

A kutatók információi hiánytalanok voltak: rendelkezésükre állt az összes létező film, valamint minden egyes bűncselekmény, amit az egész Egyesült Államokban elkövettek – méghozzá óránkénti bontásban, ami fontosnak bizonyult.

A tanulmány szempontjából kulcsfontosságú volt, hogy egyes hétvégeken erőszakos filmek vitték a prímet (például a *Hannibal* vagy a *Holtak hajnala*), máskor pedig erőszakmentesek, például az *Oltári nő* vagy a *Toy Story*.

A két közgazdász láthatta, hogy pontosan hány gyilkosságot, nemi erőszakot, testi sértést követnek el azokon a hétvégeken, amikor a mozikba kerül egy-egy erőszakos film, és összevethették a számokat azon hétvégék adataival, amikor erőszakmentes film premierjére került sor.

Mi derült ki tehát? Vajon nőtt a bűnözési ráta az erőszakos filmek bemutatása után, ahogy egyes kísérleti eredmények sugallják? Vagy változatlan maradt?

A kutatók arra jutottak, hogy a népszerű erőszakos filmek premierje után a bűnözés visszaesett.

Jól olvasták. Azokon a hétvégéken, amikor az amerikaiak milliói egymást gyilkoló embereket láttak a mozivászonon, visszaesett a bűnözési ráta, méghozzá jelentősen.

Ha az ember ennyire furcsa és ennyire váratlan eredményre jut, akkor az első gondolata az, hogy valamit rosszul csinált. A szerzők tehát külön-külön, gondosan átnézték a kódolást, de abban nem volt hiba. Az ember második gondolta ilyenkor az, hogy valami más változó magyarázhatja az eredményt. Megnézték tehát, nem befolyásolta-e az adatokat, hogy melyik évszakban vetítették a filmet. Nem befolyásolta. Ezután időjárási adatokat gyűjtöttek, hátha ez áll a háttérben. De nem.

„Ellenőriztük minden egyes feltevéseünket, mindent ellenőriztünk, amit csináltunk” – mesélte nekem Gordon Dahl. „De nem találtunk hibát.”

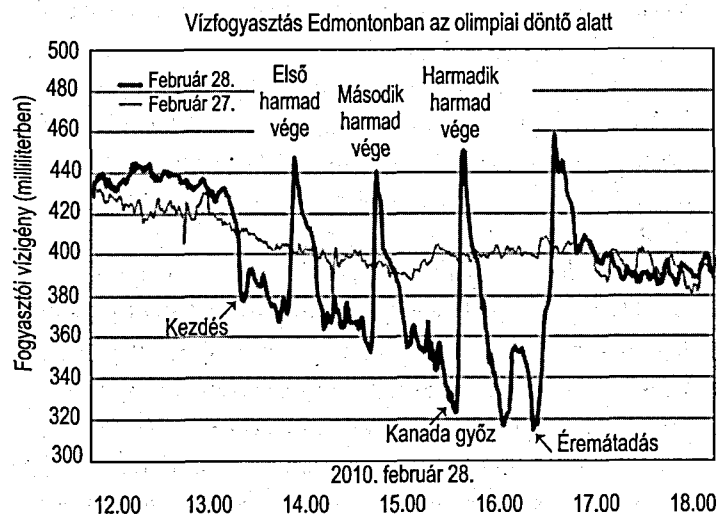
Tehát hiába sugallták az ellenkezőjét mind az anekdotikus, mind pedig a laboratóriumi bizonyítékok, hiába tűnt végtelenül különösnek az eredmény, az erőszakos filmek premierje valamiért csökkentette a bűnözési rátát. Hogy lehet ez?

A megoldás kulcsa az volt, hogy Dahl és DellaVigna a Big Data segítségével jobban rá tudott közelíteni a vizsgált adatokra. A hagyományos felmérések csak évenkénti, legfeljebb havonkénti adatokkal tudnak szolgálni. Ha szerencsénk van, egy hétvége adatait is megszerezhetjük. Manapság viszont egyre szélesebb körű, egyre átfogóbb adathalmazokból dolgozunk, nem csak kis létszámú felmérésekből, így az óránkénti, sőt a percenkénti adatokat

is tudjuk vizsgálni. Ezáltal pedig sokkal többet tudhatunk meg az emberi viselkedésről.

Az idő függvényében mért fluktuáció olykor mókás, bár nem igazán falrengető felismeréseket eredményez. Az EPCOR nevű edmontoni közművállalat beszámolt a 2010-es téli olimpia jéghokidöntője (Egyesült Államok-Kanada) alatti vízfogyasztás percenkénti alakulásáról. Az adataik szerint a fogyasztás mindig a harmad vége után ugrott meg hirtelen. Világos volt, hogy ezekben a percekben Edmonton-szerte bőszen öblítettek a vécék.

A Google-kereséseket is ki lehet gyűjteni percenkénti bontásban, ami érdekes mintázatokra derít fényt. A „blokkolatlan játékokra” irányuló keresések például hétköznapi reggel 8-kor tetőznek, és egészen délután 3-ig magas szinten maradnak. Ez nyilván az iskolák azon próbálkozására adott reakció, hogy nem veszik el a diákoktól a telefont, de blokkolják a mobilon játszható játékokat.



Az „időjárás”, „ima” és „hírek” keresőszavak hajnali fél hat előtt tetőznek, ami azt igazolja, hogy az emberek többsége sokkal korábban kel, mint én. Az „öngyilkosság” keresőszó éjjel 0 óra 36-kor ér a csúcra, és reggel 9 óra körül írják be a legkevesebben. Ez azt igazolja, hogy az emberek többsége nem érzi magát olyan nyomorultul reggelenté, mint én.

Az adatok tanúsága szerint hajnali 2 és 4 óra között jön el a nagy kérdések ideje: az olyanoké, mint a „Mi a tudat?”, a „Létezik-e szabad akarat?” vagy a „Van-e élet más bolygón?”. Az efféle kérdések éjszakai népszerűsége talán részben a cannabishasználatnak tudható be. A „hogya kell jointot sodorni?” keresőkifejezés éjjel 1 és 2 óra között tetőzik.

Dahl és DellaVigna pedig azt tudta megnézni az óriási adathalmaz segítségével, hogy miként alakult az óránkénti bűnözési ráta azokon a bizonyos mozihétfőeken. Kiderült, hogy az olyan napokon, amikor népszerű erőszakos filmet mutattak be a mozik, kora estétől lett alacsonyabb a bűncselekmények száma más hétfőké adataival összevetve. Azaz már akkor is alacsony volt a bűnözési ráta, amikor még nem peregtek az erőszakos képsorok, hiszen a nézők épp csak elfoglalták a helyüket.

Kitalálják, miért? Gondoljanak először is arra, hogy vajon kik választanak jó eséllyel hétfői programnak egy erőszakos filmet! Nyilván a fiatal férfiak – különösen az agresszív fiatal férfiak.

Most pedig gondoljanak arra, hogy hol követik el jellemzően a bűncselekményeket! Moziban, ugye, elég ritkán történik ilyesmi. Kivételek persze vannak, a legemlékezetesebb talán a 2012-ben végrehajtott, előre kitervelt colorádói lövöldözés volt. De összességében elmondható, hogy aki moziba megy, az fegyvertelen, és csöndben üldögél a helyén.

Ha az agresszív, fiatal férfiaknak felkínáljuk a lehetőséget, hogy nézzék meg a *Hannibalt*, rohannak a moziba. Ha viszont azt kínáljuk fel, hogy nézzék meg az *Oltári nőt*, akkor ezt alighanem inkább kihagyják, és helyette elmennek kocsmába, szórakozóhelyre, biliárdterembe: vagyis olyan helyre, ahol gyakoribbak az erőszakos bűncselekmények.

Az erőszakos filmek távol tartják a potenciálisan erőszakos embereket az utcáktól.

Rejtvény megoldva. Igaz? Azért még nem egészen. Az adatok ugyanis fényt derítettek még egy furcsaságra: a bűnt megelőző hatás már a vetítés elején érződött, de akkor sem szűnt meg, amikor már véget ért a film, és bezárt a mozi. Az olyan estéken, amikor erőszakos filmet mutattak be, éjjeltől reggel hatig folyamatosan alacsonyabb szinten maradt a bűnözési ráta, mint máskor.

Még ha addig, amíg azok a bizonyos fiatal férfiak bent ültek a moziban, alacsonyabb volt is a bűnözési ráta, nem kellett volna megugrania a film után, amikor már nem voltak lefoglalva? Épp végignézték egy erőszakos filmet, amiről a kísérletek bebizonyították, hogy az ember indulatosabb és agresszívebb lesz tőle.

Van ötletük, hogy mivel magyarázható az ellentmondás? A kutatóknak hosszas fejtörés után lett részük újabb „aha-élményben”. Eszükbe jutott, hogy a bűncselekmények egyik fő mozgatórugója sokszor az alkohol. És elég filmet végigültek már ahhoz, hogy tudják: az Egyesült Államokban alig-alig akad olyan mozi, ahol szeszes italt lehet kapni. Az adatok pedig igazolták, hogy az erőszakos filmek vetítése utáni, késő esti órákban meredeken zuhant az alkoholos befolyásoltság alatt elkövetett bűncselekmények száma.

Dahl és DellaVigna eredményei persze korlátozott érvényűek. Nem tudták például tesztelni a hónapokon átívelő, akár hónapokig tartó hatást – megvizsgálni, hogy meddig

tart a bűnözési ráta alacsonyabb szintje. Így hát továbbra is van rá esély, hogy az erőszakos filmeknek való tartós kitettség végső soron nem kevesebb, hanem több erőszakot eredményez. De az mindenképpen elmondható, hogy a tanulmányuk szélesebb kontextusba helyezi az említett kísérletek fő témáját, az erőszakos filmek azonnali hatását. Lehetséges, hogy vannak, akiket egy erőszakos film valóban befolyásol, és szokatlan indulatot, agresszivitást vált ki belőlük. Ugyanakkor tudják, mi az, ami egészen biztosan az erőszak irányába tolja az embereket? Az, ha más, potenciálisan erőszakos emberekkel lógnak együtt, és isznak.*

Ennek így már van értelme. Nem volt azonban értelme egészen addig, amíg Dahl és DellaVigna hozzá nem fogott volna egy nagy rakás adat elemzéséhez.

Van még egy fontos dolog, amely mindjárt világossá válik, amint rázoomolunk az adatok egy részére: az, hogy a világ bonyolult. Amit ma teszünk, annak hosszú távú hatásai lehetnek, többnyire akaratainkon kívül. A gondolatok

* A sztori azt példázza, hogy a rossznak tűnő dolgok jók is lehetnek arra, hogy megelőzzenek valami még rosszabbat. Ed McCaffrey, a Stanfordon végzett egykori elkapójátékos így magyarázta el a FoxSports csatornán, a *The Herd with Colin Cowherd* című műsorban, hogy miért hagyja amerikai futballozni mind a négy fiát: „A srácokban túlteng az energia. Ha éppen nem futballoznak, akkor gördeszkáznak, fára másznak, fogócskáznak az udvaron, vagy paintballoznak. Olyan nincs, hogy csak ülnek, és nem csinálnak semmit. Ezért én úgy vagyok vele, hogy a futball legalább egy sport, amelynek szabályai vannak. A srácaim jó párszor megfordultak már a balesetnél: lezúgtak a verandáról, elzúgtak bringával, a gördeszkával, leestek a fáról, meg ami kell... Igen, a futballban kemény ütközések vannak. De a srácaimnak olyan a természetük, hogy ez belefér. Így legalább nem szikláról ugrálnak. Mert a futball nem efféle örültség, hanem afféle kontrollált agresszió”. McCaffrey érvelése akkor az újdonság erejével hatott rám-interjú a műsorban”, de miután elolvastam Dahl és DellaVigna tanulmányát, kezdtem komolyan venni. A terepről származó, óriás adathalmazoknak megvan az az előnyük a laboradatokkal szemben, hogy kimutatják az efféle hatásokat.

is terjednek, hol lassan, hol pedig exponenciálisan, mint a vírusok. Az emberek kiszámíthatatlanul reagálnak az összefonódó hatásokra.

Az efféle összefüggéseket, kapcsolatokat, hullámhegyeket és hullámvölgyeket nem lehet kinyomozni kis létszámú felmérésekkel, illetve hagyományos adatfeldolgozó módszerekkel. Egyszerűen túl komplex, gazdag és változatos a világ ahhoz, hogy a kis adat labdába rúghasson itt.

ALTEREGŐINK

2009 júniusában úgy nézett ki, hogy David „Big Papi” Ortiznak annyi. Pedig az azt megelőző fél évtized remekül sikerült: a bostoniak szabályosan beleszerettek a barátságos mosolyú, szellős fogsorú, dominikai születésű sztárfutballistába.

Ortiz ez idő alatt zsinórban ötször került be az All Star-csapatba, egyszer elnyerte a liga legértékesebb játékosának díját, és nagyban hozzájárult ahhoz, hogy a Boston Red Sox – nyolcvanhat éves „böjt” után – újra bajnok lett. A 2008-as idényben azonban a harminckét éves Ortiz mutatói leromlottak: az ütőátlaga 68 ponttal, az első bázisra jutási átlaga 76 ponttal, a súlyozott ütőátlaga 114 ponttal lett gyengébb. A 2009-es évad elején pedig „Big Papi” számai még tovább zuhantak.

Bill Simmons sportszakíró, aki szenvedélyes Red Sox-szurkoló is egyben, így írta le, hogy mi történt a 2009-es évad első hónapjaiban: „Egyértelmű, hogy David Ortiz többé már nem kiemelkedő baseballjátékos. A nagydarab sztárfutballisták úgy működnek, mint a pornósztárok, a pankrátorok, az NBA-centerek és a »trófeafeleségek«: ha egyszer vége van, akkor vége van.” A vérbeli sportrajongó hisz a sze-

mének, és Simmons szeme most azt súgta, hogy Ortiznak „vége van”. Konkrétan azt jósolta, hogy hamarosan kispadra ültetik, vagy megválnak tőle.

De vajon tényleg vége volt Ortiznak? Ha ön lenne a 2009-es Boston Red Sox menedzsere, kitenne a csapatból? Az általánosabb érvényű kérdés pedig így szól: miként jósolhatjuk meg egy baseballjátékos jövőbeni teljesítményét? A még általánosabb érvényű kérdés pedig így: miként jósolhatjuk meg a Big Data segítségével, hogy mit fognak csinálni a jövőben az emberek?

Messzire juthat az adattudományban, aki magáévá teszi az alábbi elméletet: nézzük meg, mit csinálnak a baseball-adatok tudorai, a „szabermetrikusok” (a szó az Amerikai Baseballkutató Társaság rövidített nevéből, a SABR-ból ered), és számítsunk rá, hogy a módszereik futótűzként fognak terjedni! A baseball az első olyan szakterületek egyike volt, ahol mindenre kiterjedő adathalmazok álltak rendelkezésre nagyjából mindenről, és ahol okos emberek egész hada tette fel az életét arra, hogy kisüssön valamit a temérdek adatból. Ma már szinte minden tudományág itt, vagy legalábbis érrefelé tart. A baseball halad az élen, a többi tudományág pedig követi a példáját. A „szabermetrika” megeszi őket reggelire.

Egy baseballjátékos jövőjét úgy jósolhatjuk meg a legegyszerűbben, ha abból indulunk ki, hogy továbbra is úgy fog játszani, ahogy most játszik. Ha tehát valaki az utóbbi másfél évben csak kínlódott, akkor arra tippelhetünk, hogy a következő másfél évben is kínlódni fog.

E módszertan szerint a Boston Red Soxnak ki kellett volna raknia a csapatból David Ortizt.

Akadhatnak azonban más releváns információk is. Bill James, akit a legtöbben a „szabermetrika” alapító atyjának tartanak, még az 1980-as években erősen hangsúlyozta

az életkor jelentőségét. Ő úgy találta, hogy a baseballjátékosok hamar, már 27 éves koruk körül elérnek a csúcra. A klubok viszont hajlamosak voltak figyelmen kívül hagyni, mennyit romlik a játékosok teljesítménye az évek előrehaladtával, és túlfizették korosodó sztárjaikat.

Ha a Boston Red Sox ezzel a korábbinál fejlettebb módszerrel dolgozik, egyértelműen ki kellett volna raknia a csapatból David Ortizt.

Csakhogy még így is mellétrafálhatunk, hiszen nem minden játékos járja be ugyanazt az utat: akad, aki 23 évesen teljeseodik ki, és olyan is van, aki 32 évesen. Továbbá nem ugyanúgy öregszenek a magas, az alacsony, a kövér, a sovány játékosok. A baseballstatisztikusok kimutatták, hogy a különböző típusú játékosoknak a pályáíve is különbözik. De ez csak rontott Ortiz helyzetén, hiszen a nagydaráb sztárfutballisták valóban korán eljutnak a csúcra, majd harminc után nagyon gyorsan hanyatlani kezdenek.

A Boston Red Soxnak tehát a közelmúltban nyújtott teljesítménye, a kora és a méretei alapján kétségtelenül ki kellett volna raknia a csapatból David Ortizt.

Csakhogy Nate Silver, a statisztikus 2003-ban kidolgozott egy új modellt, amelyet úgy hívott, hogy PECOTA. Ez pedig a jövőmondás legjobb, egyúttal legvagányabb módszerének bizonyult. Silver felkutatta minden egyes játékos alteregóját. Elmondom, hogyan működik a dolog. Először is rakjuk össze a Major League Baseball teljes körű adatbázisát: legyen benne mindenki, aki valaha is játszott a ligában. És legyen benne minden, amit csak tudunk erről a több mint 18 ezer emberről: a magasságuk, az életkoruk, a posztjuk, az ütőátlaguk, valamint a hazafutásaik, a sétáik és az ütőhiba miatti kieséseik száma, mindez a pályafutásuk éves bontásában. Ha ez megvan, akkor megkereshetjük azt a húsz játékost, akik a karrierjük aktuális szaka-

szában a legjobban hasonlítottak Ortizra – vagyis azokat, akik úgy játszottak, mint ő 24, 25, 26, 27, 28, 29, 30, 31, 32, illetve 33 éves korában. Másképp fogalmazva: keressük meg Ortiz alteregóit, és nézzük meg, hogy alakult a későbbi pályafutásuk!

Az alteregókeresés újabb példa a zoomolásra. Ez a módszer ráközelít az adott személyre leginkább hasonlító személyek apró részalmazára. És, ahogy ez a zoomolásnál lenni szokott, minél több adatunk van, annál jobb. Ortiz esetében például kiderült, hogy az alteregói egészen más jövőt vetítenek előre számára, mint a többi módszer. Ortiz alteregói között ott találjuk Jorge Posadát és Jim Thome-ot. Ezeknek a játékosoknak kissé lassan indult be a karrierjük: a húszas éveik végén berobbantak az élvonalba, és világklasszissá nőttek ki magukat; a harmincas éveik elején pedig csak vergődtek.

Silver annak alapján jósolta meg Ortiz későbbi teljesítményét, hogy mi történt ezt követően az alteregóival. Hát ez: Posada és Thome visszanyerte az erejét. A „trófeafeleiségekre” igaz lehet, amit Bill Simmons mondott: ha egyszer vége van, akkor vége van. Ortiz hasonmásainál viszont más volt a helyzet: ami egyszer elment, az később mégiscsak visszajött.

A teljesítmény előrejelzésére valaha alkalmazott legjobb módszer, az alteregókeresés tehát azt hozta ki, hogy a bostoniak legyenek türelmesek Ortizzal. És a Red Sox valóban türelmes volt a korosodó sztárfutballistával. 2010-re aztán Ortiz ütoátlaga elérte a .270-et, vagyis a 27%-ot. Ortiz abban az idényben 32 hazafutást ütött, és bekerült az All Star-csapatba, majd az ezt következő három idényben újra. 2013-ban aztán a 37 éves Ortiz a hagyományos posztján, harmadik ütőjátékosként .688-as ütőátlagot produkált a bajnoki döntő hat meccsén, amikor is a Red Sox 4:2-re

verte a St. Louist. Ortiznak meg is szavazták a World Series MVP Awardot, a Major League döntőjének legértékesebb játékosát megillető elismerést.*

Miután befejeztem Nate Silver eszmefuttatását a játékosok pályáivét előre jelző módszeréről, egyből szöveget ütött a fejemben, hogy talán nekem is lehet alteregóm.

Az alteregókeresés nem csak a sportban tartogat ígéretes lehetőségeket. Nyomára bukkanhatok például annak a személynek, akivel a leginkább egyezik az érdeklődési körünk? Ha megtalálnám, talán lóghatnánk együtt. Talán tudna pár jó éttermet, és mutathatna nekem egy-két dolgot, amihez van affinitásom, csak eddig nem tudtam róla.

Az alteregókeresés egyénekre, sőt azok egyéni jellemvonásaira közelít rá. És ahogy az a zoomolásnál lenni szokott, minél több adatunk van, annál élesebb lesz a kép. Tegyük fel, hogy egy tíz emberből álló adathalmazban keresem a hasonmásomat! Talán rábukkanok valakire, aki szeret olvasni, akárcsak én. Tegyük fel, hogy egy nagyjából ezer emberből álló adathalmazban keresem a hasonmásomat! Talán rábukkanok valakire, akinek bejönnek az ismeretterjesztő jellegű fizikakönyvek. De tegyük fel, hogy egy több százmillió emberből álló adathalmazban keresem a hasonmásomat. Ebben talán már olyat is találhatok, aki tényleg, komolyan hasonlít rám.

Egy szép napon alteregóvadászatra indultam a közösségi oldalakon. A Twitter-profilok teljes adatállományában

* Azoknak, akik eljutottak idáig a könyvben, már valószínűleg feltűnt, hogy hajlamos vagyok cinikusan viszonyulni a szép történetekhez. Ezáltal azonban azt akartam, hogy egy sztori végre jó véget érjen, ezért a lábjegyzetbe számúztam a cinizmusomat. Élek a gyanúperrel, hogy a Nate Silver-féle PECOTA valójában csak a következőkre derített fényt: Ortiz koksolt, majd leállt vele, de megint el fogja kezdeni. A teljesítményjóslás szempontjából így is tök jó, hogy a PECOTA ezt kiszúrta, de így már nem annyira megindító a történet.

kerestem bolygónk azon lakóit, akiknek érdeklődési köre a legtöbb átfedést mutatja az enyémmel.

Az érdeklődési körömről nyilván igen sokat elárul, hogy kiket követek a Twitteren. Összesen úgy 250 emberről van szó, és a névsorból kiderül, hogy szenvedélyesen érdekel a sport, a politika, a humor, a természettudomány és a mogorva zsidó folkénekések.

Van-e tehát az univerzumban még valaki, aki ugyanazt a 250 embert követi, létezik-e a Twitter-ikertestvérem? Természetesen nem. A hasonmásaink nem azonosak velünk, csak hasonlítanak ránk. Olyan ember sem létezik, aki 200 Twitter-fiókot követne azok közül, amelyeket én követek. Vagy akár csak 150-et.

Mindazonáltal legvégül rábukkantam egy olyan Twitter-fiókra, amely elképesztő módon százat követett az én 250-es listámból: úgy hívták, hogy Country Music Radio Today. Jó, mi? De mint kiderült, a Country Music Radio Today egy botprogram volt (már nem létezik), amely 750 ezer Twitter-fiókot követett abban a reményben, hogy a közeledése viszonzásra talál.

Az egyik exbarátnőm, akiről gyanítom, hogy padlót fogna a röhögéstől, ha ezt megtudná. Egyszer azt mondta nekem, hogy inkább vagyok robot, mint ember.

De a viccet félretéve: az első kutatási eredményem, amely szerint az alteregóm egy 750 ezer random Twitter-profil követő botprogram, elárul valami fontosat az alteregókérésről. A módszer akkor ér valamit, ha nemcsak azt keressük, hogy ki kedveli ugyanazt, amit mi, hanem azt is, hogy ki nem kedveli ugyanazt, amit mi.

Hogy engem mi érdekel, az nemcsak abból derül ki, hogy milyen profilekat követek, hanem abból is, hogy mi az, amit inkább nem követek. Érdekel, ugyebár, a sport, a politika, a humor és a természettudomány, nem érdekel

viszont az étel, a divat és a színház. A követéseim azt mutatják, hogy Bernie Sanderst kedvelem, Elizabeth Warrent nem, Sarah Silvermant igen, Amy Schumert nem, a *New Yorkert* igen, az *Atlanticet* nem, a barátaim közül Noah Poppot, Emily Sandset és Josh Gottliebét igen, Sam Ashert viszont nem. (Bocs, Sam, de a Twitter-hírfolyamodon bealszik az ember.)

Kinek a profilja hasonlít tehát a legjobban az enyémhez a 200 millió twitterező közül? Kiderült, hogy az alteregóm a *Vox* hírportál munkatársa, Dylan Matthews. Ez kissé lelelombozott, hiszen ez nem segíti elő médiafogyasztási szokásaim színesítését: Matthewst ugyanis már követem a Twitteren és a Facebookon, és mohón falom a *Vox*-posztjait. Így hát nem igazán változtatta meg az életemet a hír, hogy ő az alteregóm. Persze így is elég menő, hogy tudom, ki hasonlít rám a legjobban a világon, pláne úgy, hogy rajongok az illetőért. Ha befejezem a könyvet és felhagyok a remetéélettel, talán összefuthatunk Matthews-szal, és elbeszélgethetünk James Surowiecki írásairól.

A baseballrajongók biztosan élvezték az Ortiz-hasonmás utáni kutatást. A saját alteregóm keresése is meglehetősen szórakoztató volt, legalábbis nekem. De vajon mit árulnak el még az efféle keresések? Hogy csak egy példát mondjak, a nagy internetes cégek közül jó néhány él ezzel az alteregókéréssel annak érdekében, hogy elképesztő ütemben fejlesszék a kínálatukat és a felhasználói élményt. Az Amazon például az alteregókéréshez hasonló módszerrel dönti el, hogy kinek milyen könyveket ajánljon. Megnézik, mit választanak az adott felhasználóhoz hasonló emberek.

A Pandora is így választja ki, hogy milyen dalokat hallgatnánk meg szívesen. A Netflix is innen következteti ki, hogy milyen filmeket szeretnénk megnézni. Az olvasói

preferenciák alteregókereséssel történő előrejelzése igen csak bevált: amikor Greg Linden, az Amazon informatikusa bevezette a módszert, a könyvajánlások hatékonysága annyit javult, hogy Jeff Bezos, az Amazon alapítója térdre borult Linden előtt, és így kiáltott: „Nem vagyok méltó!”

De a módszer hatékonyságának fényében nem is az a lényeg, hogy mennyire elterjedt az alteregókeresés: inkább az, hogy milyen sokszor nem alkalmazzák. Elég fontos területei vannak az életnek, ahol ugrásszerű fejlődést eredményezhetne az a fajta személyre szabás, amelyet ez a módszer lehetővé tenné. Jó példa erre az egészségügy.

Isaac Kohane, aki informatikával és orvosi kutatásokkal foglalkozik a Harvardon, az orvoslásba próbálja bevonni az alteregókeresést. Úgy próbálja megszerezni és gyűjteni az egészségügyi adatokat, hogy az orvos ne ugyanazt a sémát húzza rá mindenre, hanem meg tudja keresni az épp aktuális pácienséhez hasonló eseteket. Ezzel a módszerrel a diagnosztika, illetve a terápia sokkal inkább személyre szabottá válhatna.

Kohane szerint ez a fajta szemlélet magától értetődő az orvoslásban, vagyis még csak nem is lenne ez valami radikális lépés. „Hiszen mit állít egy diagnózis?” – teszi fel a kérdést Kohane. „Azt, hogy közös jellemzőid vannak egy korábban vizsgált populációval. Ha például, isten ne adja, infarktussal diagnosztizálalak, azzal azt mondom, hogy a patofiziológiád szerint, amelyet más embereknél már látam, infarktusod van.”

A diagnózis tehát egyfajta kezdetleges alteregókeresés. A gond csak az, hogy az orvosok által használt adatbázisok nagyon kicsik. Egy orvos azon páciensek populációjára alapozza a diagnózist, akiket addig kezel, illetve esetleg a szaklapokból megismert, szintén kis létszámú populációkra, amelyeket a kutatók vizsgáltak. Csakhogy,

mint láttuk, az alteregókeresés akkor működhet igazán jól, ha az adathalmaz sok-sok további esetet is tartalmaz.

Ezen a területen tehát sokat segíthetne egy kis Big Data. Akkor hát mi tart ennyi ideig? Miért nem alkalmazzák máris széles körben? A probléma az adatgyűjtésben rejlik. A kórtörténetek zöme ugyanis mindmáig csak papíron létezik, vagy ha számítógépre viszik, akkor is sokszor elzárják valami inkompatibilis formátumba. Amint azt Kohane megjegyezte, a baseballról sokszor jobb minőségű adataink vannak, mint az egészségünkről. Pedig már egy-két pofonegyszerű intézkedés is komoly eredményt hozhatna. Kohane sokszor használja az „alacsony ágról lelógó gyümölcs” kifejezést. Szerinte már az is forradalmasíthatná a gyermekgyógyászatot, ha sikerülne létrehozni a magasság- és testsúlygrafikonok teljes körű adathalmazát. Így össze lehetne hasonlítani a gyerekek növekedésének ütemét: a számítógép kidobná a hasonló tempóban fejlődött ifjakat, és automatikusan kidobná az aggasztó jeleket. Észlelné, ha egy gyerek idő előtt megáll a növésben, aminek bizonyos körülmények között két oka lehet: pajzsmirigy-alulműködés vagy agydaganat. Mindkét esetben nagy áldás a korai diagnózis. Ismét Kohane-t idézem: „Ez a két betegség ritka madár, úgy tízezerből egyszer fordulnak elő. A gyerekek nagy többsége egészséges. Úgyhogy szerintem legalább egy évvel hamarabb meg tudnánk állapítani a diagnózist. Igen, száz százalék, hogy meg tudnánk!”

James Heywood vállalkozó másként közelít az orvosi adatok összekapcsolásának nehézségeihez. Létrehozta a PatientsLikeMe.com weboldalt, ahova ki-ki a saját adatait töltheti fel: betegség, kezelés, mellékhatások. Heywood máris sikereket könyvelhet el: feltérképezi a lehetséges forogatókönyveket, és az eredményei összevethetők azzal, amit általában az adott betegségről gondolunk.

Az a célja, hogy annyi embert toborozzon és annyi betegségét fedjen le az adatbázisa, amennyi már elég ahhoz, hogy az emberek megtalálhassák a saját egészségügyi altergójukat. Heywood reményei szerint mindenki megtalálja majd a megfelelő korú és nemű, hasonló kórtörténetű, hasonló tünetekről beszámoló társait, és megnézheti, mi vált be náluk. Ez már valóban egészen másféle orvoslás lenne.

ADATSZTORIK

A zoomolás, vagyis a ráközelítés sok szempontból nagyobb értéket képvisel a szememben, mint egy adott tanulmány által feltárt konkrét tények, mivel a zoomolásnak köszönhetően másként szemlélem az életet, és másként is beszélek róla.

Amikor az emberek megtudják, hogy adattudós és író vagyok, olykor elmesélnek nekem egy-egy esetet vagy felmérési eredményt. De az én szememben az ilyen adatok sokszor unalmasak: statikusak, élettelenek. Nincs sztorijuk.

A barátaim pedig időnként arra próbálnak rávenni, hogy olvassak én is regényeket és életrajzokat. De ezek is kevésbé érdekelnek. Akaratlanul is mindig azt kérdezem magamtól: „Történhet ilyesmi más helyzetekben is? Hol van itt az általánosíthatóság?” A könyvekben elmesélt történetek aprónak és nem reprezentatívnek tűnnek.

Ezért a saját könyvemben olyasmit igyekszem bemutatni, ami szerintem semmi máshoz nem fogható. Egy adatokon és számokon alapuló módszert, amelynek igen nagy a szemléltető ereje, és meglehetősen mélyenszántó. Az adatok jelentéstartalma olyan gazdag, hogy szinte látjuk a mögöttük megbúvó embereket. Ha rázoomolunk Edmonton városának percenkénti vízfogyasztására, *látjuk magunk*

előtt a hokimeccset néző embereket, amint a harmad végén felállnak a kanapéről. Ha rázoomolunk a Philadelphiából Miamiba költöző, majd ott az adócsalásra rákapó emberekre, *látjuk magunk előtt*, ahogy a szomszédaikkal trécselve eltanulják a 9000 dolláros trükköt. Ha a különböző életkorú baseballrajongókra közelítünk rá, *látjuk magunk előtt* a gyerekkoromat és az öcsém gyerekkorát, továbbá felnőtt férfiak millióit, akik még ma is sírva fakadnak, amikor az a csapat veszít, amely nyolcéves korukban rabul ejtette a szívüket.

Újfent megkockáztatnék egy grandiózus kijelentést: a könyvemben felbukkanó közgazdászok és adattudósok nem csupán egy új eszközt hoznak létre, hanem új műfaj alapjait vetik meg. E fejezetben és a könyv nagy részében azt próbáltam bemutatni, hogy erre az irdatlan méretű, változatos, részletgazdag adattömegre nagyon jól rá lehet közelíteni. Úgy mesélhetünk el bonyolult és tanulságokkal teli történeteket, hogy nem kell egyetlen konkrét, nem reprezentatív emberre szorítkoznunk.

Labor az egész világ

2000. február 27-e átlagos napnak indult a Google Mountain View-i épületegyüttesében. Sütött a nap, tekertek a bringások, masszíroztak a masszözök, uborkalével hidralták magukat a dolgozók. De aztán ezen az átlagos napon a Google néhány mérnökének támadt egy ötlete, és rájöttek valamire, ami mára az internet egyik fő hajtóerejévé vált. Kitalálták, miként érhetik el, hogy a felhasználók a lehető legtöbbször rákattintsanak a weboldalaikra, illetve később visszatérjenek, és ott is maradjanak.

Mielőtt elmondanám, hogy pontosan mit csináltak, beszélnünk kell a korreláció kontra okság problematikájáról, amely roppant komoly kérdés az adatelemzésben, de eddig nem foglalkoztunk vele súlyának megfelelően.

Úgy tűnhet, hogy a média nap mint nap bombáz minket korrelációkon alapuló tanulmányok eredményeivel. Megtudhattuk például, hogy aki rendszeresen fogyaszt egy kevés alkoholt, az jellemzően jobb egészségnek örvend. Ez egy korreláció.

De vajon igaz-e, hogy a mérsékelt alkoholfogyasztás egészségjavító hatású, azaz fennáll-e az oksági kapcsolat? Ez már egyáltalán nem biztos. Elképzelhető, hogy épp fordítva van: az emberek azért isznak egy kicsit, mert jó egészségnek örvendenek. Ezt a társadalomtudósok fordított oksági viszonyinak nevezik. Továbbá az is elképzelhető, hogy

van egy független tényező, amely kiváltó oka mind a mérsékelt alkoholfogyasztásnak, mind a jó egészségnek. Lehet, hogy ha az ember sok időt tölt a barátaival, az egyszerre vezet mérsékelt alkoholfogyasztáshoz és jó egészséghoz. A társadalomtudósok ezt úgy hívják, hogy kihagyott változó okozta torzítás.

Hogy állapíthatjuk meg biztonsággal az oksági összefüggést? Az arany szabály az, hogy végezzünk el egy randomizált kontrollált kísérletet. A dolog úgy működik, hogy véletlenszerűen (random módon) csoportra osztjuk az embereket: a kísérleti csoport tagjainak csinálniuk vagy szedniük kell valamit, a kontrollcsoport tagjainak pedig nem. Megfigyeljük, miként reagálnak a két csoport tagjai. A két csoport eredményei közötti különbség lesz az oksági hatás.

Ha például azt akarjuk tesztelni, hogy jó egészséget okoz-e a mérsékelt alkoholfogyasztás, akkor random módon kiválaszthatunk néhány embert arra a feladatra, hogy egy éven keresztül minden nap igyanak napi egy pohár bort, a másik random csoportot pedig megkérhetjük, hogy egy éven keresztül ne igyanak egy kortyot sem. Egy évvel később pedig összevethetjük az alanyok saját bevallásuk szerinti egészségi állapotát. Mivel random módon osztottuk be őket a két csoportba, nincs okunk feltételezni, hogy valamelyik csoport eleve jobb egészségnek örvend vagy esetleg élénkebb társasági életet él – vagyis bízhatunk benne, hogy a bor hatása oksági tényező. A randomizált kontrollált kísérlet minden tudományterületen a legmegbízhatóbb módszernek számít. Ha egy gyógyszer kiállja a randomizált kontrollált kísérlet próbáját, akkor megkezdődhet a lakossági forgalmazása. Ha viszont nem megy át ezen a teszten, akkor nem kerül a patikákba.

A társadalomtudomány is egyre nagyobb mértékben alkalmazza a randomizált kísérleteket. Esther Duflo, a MIT

francia közgazdász, síkra szállt azért, hogy a fejlődésgazdaságtan is éljen ezzel a módszerrel, amely arra keresi a választ, hogyan lehetne segíteni a világ legszegényebbjein. Duflo és kollégái egyik tanulmányukban azt vizsgálták, hogy miként lehetne fejleszteni az oktatást India vidékies részein, ahol a felsősök több mint fele nem tud elolvasni egy egyszerű mondatot. Ennek egyik lehetséges oka az, hogy a tanárok nem járnak be rendszeresen órát tartani: vannak iskolák, ahol egy adott napon több mint 40 százaléuk nem jelenik meg.

Esther Duflo és munkatársai véletlenszerűen két csoportra osztották az iskolákat. Az egyikben (a kísérleti csoportban) a tanárok az alapfizetésükön felül szerény napi-díjat (50 rúpiát, azaz körülbelül 1 dollár 15 centet) kaptak, ha bementek tanítani. A másik csoportban nem járt pluszpénz az iskolában való megjelenésért. A kísérlet figyelemre méltó eredményt hozott. Ott, ahol a tanárok napi-díjat kaptak, felére csökkent az elmaradt órák száma. Jelentősen javultak a diákok teszteredményei is, különösen a kislányok esetében. A napi-díjat fizetős iskolákba járó lányok hét százalékkal jobb eséllyel tanultak meg írni a kísérlet időtartamának végére.

A *New Yorker* cikke szerint Bill Gatest annyira lenyűgözték Esther Duflo eredményei, hogy azt mondta neki: „Támogatnunk kell magát!”

AZ A/B-TESZTELÉS ABÉCÉJE

Az oksági viszony bizonyításának aranystandardja tehát a randomizált kísérlet, amelynek alkalmazása már a társadalomtudományokban is elterjedt. Most pedig kanyarodjunk vissza oda, hogy mi is történt a Google központjában

2000. február 27-én! Mivel forradalmasították ezen a napon az internetet?

Az történt, hogy néhány mérnök elhatározta: elvégeznek a Google weboldalán egy kísérletet. Véletlenszerűen két csoportra osztották a felhasználókat: a kísérleti csoportnak húsz linket mutattak a keresési eredményeket listázó oldalon, a kontrollcsoportnak pedig szokás szerint tízet. Ezután pedig összevetették a két csoport elégedettségét, vagyis azt, hogy milyen gyakran tértek vissza a Google-ra.

Ez lett volna a forradalom? Nem tűnik valami falrengető dolognak. Már említettem, hogy a randomizált kísérlet módszerét a gyógyszercegek és a társadalomtudósok is alkalmazzák. Miért olyan nagy durranás az, ha utánozzák őket?

Az itt a lényeg – amint azt a Google mérnökei villámgyorsan felismerték –, hogy a digitális világban végzett kísérleteknek van egy hatalmas előnyük az offline kísérletekkel szemben. A randomizált offline kísérletek meggyőzőek ugyan, de rendkívül erőforrás-igényesek is. Ahhoz, hogy Esther Duflo tanulmánya elkészülhessen, iskolákkal kellett kapcsolatba lépni, meg kellett szervezni a finanszírozást, fizetni kellett a tanárok egy részét, és tesztelni kellett az összes diákot. Az offline kísérletek több ezer, vagy akár több százezer dollárba is kerülhetnek, és hónapokig, évekig is eltarthatnak.

A digitális világban viszont a randomizált kísérlet gyors és olcsó. A résztvevőket sem toborozni, sem fizetni nem kell. Ehelyett egyszerűen írunk egy kódsort, amellyel random módon hozzárendeljük őket valamelyik csoporthoz. A felhasználóknak kérdőívet sem kell kitölteniük. Egyszerűen mérhetjük az egérmozdulataikat és a kattintásaikat. A válaszaikat sem kell kézzel kódolnunk és elemeznünk, hanem írhatunk egy programot, amely ezt automatikusan

elvégez. Nem kell kapcsolatba lépniük senkivel, még csak nem is kell szólnunk a felhasználóknak, hogy egy kísérlet alanyai.

Ez a Big Data negyedik különleges képessége: nagyon, de nagyon megkönnyíti a randomizált kísérleteket, amelyek képesek az igazi oksági viszonyok feltárására. Bármikor és szinte bárhol lebonyolítható egy kísérlet, csak online kell lenni hozzá. A Big Data korában labor az egész világ.

A nagy felismerés futótűzként terjedt a Google-nál, majd az egész Szilícium-völgyben, ahol a randomizált kontrollált kísérletek az „A/B-tesztelés” nevet kapták. A Google mérnökei 2011-ben hétezer A/B-tesztet futtattak le, és ez a szám azóta egyre csak emelkedik.

Ha a Google azt szeretné elérni, hogy az emberek többször kattintsanak a weboldalaikon lévő hirdetésekre, próbálkozhat, mondjuk, a kék szín két különböző árnyalatával: az egyiket az A csoport kapja, a másikat a B csoport, majd összevetik a kattintásszámokat. Az, hogy az efféle teszteket ilyen könnyű elvégezni, természetesen túlhasználathoz vezethet: akadtak is olyan Google-alkalmazottak, akik úgy érezték, hogy a cég túlzásba viszi az erőfeszítést nem igénylő kísérletezéssel. Egy frusztrált formatervező 2009-ben fel is mondott, miután a Google az A/B-teszttel végigpróbálgatta a kék szín negyvenegy, egymástól alig megkülönböztethető árnyalatát. De az, hogy egy dizájnert többre tartotta a művészetet a rögeszmés piackutatásnál, nem gátolta a módszer terjedését.

A Facebook ma már napi ezer A/B-tesztet futtat, ami azt jelenti, hogy az ott dolgozó mérnökök kis csapata több randomizált kontrollált kísérletbe fog bele naponta, mint az egész gyógyszeripar egy teljes év alatt.

Az A/B-tesztelés többé nem a legnagyobb techcégek kiváltsága. Dan Siroker, a Google volt alkalmazottja bevonta

a módszert Barack Obama első elnökválasztási kampányába: így tesztelték a honlapjaik dizájnját, a sajtónak témát javasoló egymondatos e-maileket, az úgynevezett „pitch-eket”, valamint az adományozási űrlapokat is. Utána Siroker létrehozta új cégét, az Optimizelyt, amely szervezeteknek nyújt segítséget gyors A/B-tesztelések elvégzésében. Az Optimizely szolgáltatait 2012-ben nemcsak Obama, hanem ellenfele, Mitt Romney is igénybe vette a szavazók, az önkéntesek és az adományozók mozgósítására. Ezenkívül a legkülönbözőbb cégektől kapnak megbízást, így a Netflix-től, a TaskRabbit-től és a New York magazintól is.

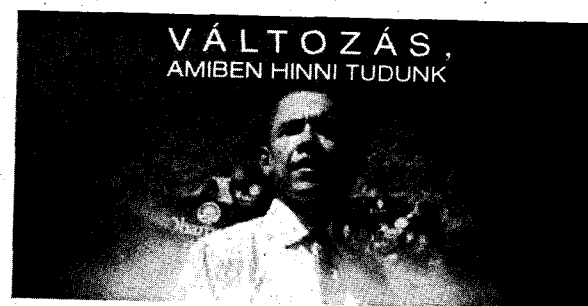
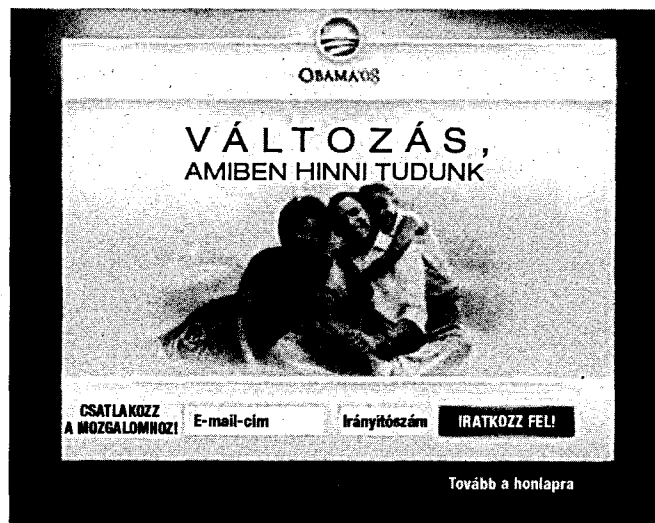
Hogy lássuk, mekkora értéket képvisel az A/B-tesztelés, nézzük meg, hogyan használta fel Obama a kampányában résztvevők létszámának növelésére. A honlapon eleinte csak az elnökjelölt képe díszelgett, alatta pedig „Sign Up” gomb biztatta az embereket a feliratkozásra.



De valóban így érdemes köszönteni az embereket a honlapon? Obama csapata Dan Siroker segítségével le tudta tesztelni, hogy esetleg több embert késztet-e feliratkozásra egy másik kép vagy egy másik gomb. Lehet, hogy többen kattintának a gombra, ha Obama komolyabb arcot vágna a képen? Vagy ha a gomb felirata inkább az lenne, hogy „Join Now”, vagyis „Csatlakozz most!”? A kampánystáb különféle kép-gomb párosításokat próbált ki: azt vizsgálták, hogy melyik hány kattintást hoz. Lássuk, kitalálja-e a kedves Olvasó, hogy melyik lett a nyerő kombináció!

A győztes az Obama-család fotója és a „Learn More” (Tudj meg többet!) gomb lett. Még hozzá hatalmas fölényrel: az Obama-kampánystáb becslése szerint 40 százalékkal többen iratkoztak fel, és nagyjából 60 millió dollár pluszadomány futott be.

Van még egy óriási előnye annak, hogy ez a jól bevált tesztelési forma ilyen olcsón és egyszerűen kivitelezhető: még inkább megszabadít minket annak terhetől, hogy az



Tesztelt gombok

CSATLAKOZZ MOST!

TUDJ MEG TÖBBET!

IRATKOZZ FEL!

intuíciónkra kelljen hagyatkoznunk, amelynek, amint arról az első fejezetben már szó esett, megvannak a maga korlátai. Az A/B-tesztelés jelentőségének egyik alapvető oka az, hogy az emberek kiszámíthatatlanok. Ezért van az, hogy az intuíciónk sokszor nem képes megjósolni a reakcióikat.

Eltalálta, milyen lett az optimális Obama-honlap?

Hadd mutassak még néhány „intuícíótesztet”! A *The Boston Globe* A/B-teszteli a szalagcímeket, hogy lássa, melyekre kattintanak rá többen. Próbálja kitalálni, hogy az alábbi párokból melyik lett a győztes!

AZ EGYIK SZALAGCÍM SOKKAL TÖBB KATTINTÁST HOZOTT,
MINT A MÁSIK

	A SZALAGCÍM	B SZALAGCÍM
1.	Megmentheti-e a bálnákat a SnotBot drón?	Megmentheti ez a drón a bálnákat?
2.	Naná hogy Massachusettsben toplistas keresőszo a „puha tojas”!	Ez a toplistas massachusettsi Google-kereses elég kinos
3.	Csajozoverseny áll a St. Paul Középiskolában történt nemi erőszak háttérében	Nincs vádemelés az elit gimnáziumban történt szexbotrányban
4.	Bankot robbantott a nő a ritka baseballkártyával	179 ezer dollárt keresett a nő a ritka baseballkártyán
5.	A bostoni közlekedési vállalat prognózisa: 2020-ig megduplázódik az éves működési veszteség	Felkészülni: duplájára nő a bostoni közlekedési vállalat hiánya!
6.	Így mozdította elő Massachusetts a fogamzásgátláshoz való hozzáférést	Így segített eltörölni a Bostoni Egyetem a „közerkölc elleni bűnöket”
7.	Amikor megnyitották Bostonban az első metróvonalat	Karikatúrák az első bostoni metróvonal megnyitásáról

	A SZALAGCÍM	B SZALAGCÍM
8.	A mételyező kultúrát okolja az elit gimnáziumban történt nemi erőszak áldozata és családja	Közleményt adott ki az elit gimnáziumban történt nemi erőszak áldozata és családja
9.	Csak egy „Szabadságot Bradynek”-sapkás fickó tudta lebuktatni a trükkös Miley Cyrust	Sok szépet láthatott a Patriots-drukker, amiért kiszúrta az álrúhás Miley Cyrust

Készen vannak a tippeléssel? Alább kövérrel szedve láthatók a helyes válaszok.

	A SZALAGCÍM	B SZALAGCÍM	GYŐZTES
1.	Megmentheti-e a bálnákat a SnotBot drón?	Megmentheti ez a drón a bálnákat?	53%-kal több kattintás az A-ra
2.	Naná hogy Massachusettsben toplistas keresőszo a „puha tojas”!	Ez a toplistas massachusettsi Google-kereses elég kinos	986%-kal több kattintás a B-re
3.	Csajozoverseny áll a St. Paul Középiskolában történt nemi erőszak háttérében	Nincs vádemelés az elit gimnáziumban történt szexbotrányban	108%-kal több kattintás a B-re
4.	Bankot robbantott a nő a ritka baseballkártyával	179 ezer dollárt keresett a nő a ritka baseballkártyán	38%-kal több kattintás az A-ra
5.	A bostoni közlekedési vállalat prognózisa: 2020-ig megduplázódik az éves működési veszteség	Felkészülni: duplájára nő a bostoni közlekedési vállalat hiánya!	62%-kal több kattintás a B-re
6.	Így mozdította elő Massachusetts a fogamzásgátláshoz való hozzáférést	Így segített eltörölni a Bostoni Egyetem a „közerkölc elleni bűnöket”	188%-kal több kattintás a B-re

	A SZALAGCÍM	B SZALAGCÍM	GYŐZTES
7.	Amikor megnyitották Bostonban az első metróvonalat	Karikatúrák az első bostoni metróvonal megnyitásáról	33%-kal több kattintás az A-ra
8.	A mételyező kultúrát okolja az elit gimnáziumban történt nemi erőszak áldozata és családja	Közleményt adott ki az elit gimnáziumban történt nemi erőszak áldozata és családja	76%-kal több kattintás a B-re
9.	Csak egy „Szabadságot Bradynek”-sapkás fickó tudta lebuktatni a trükkös Miley Cyrust	Sok szépet láthatott a Patriots-drukker, amiért kiszúrta az álház Miley Cyrust	67%-kal több kattintás a B-re

Arra tippelek, hogy olvasóim a helyes válaszok több mint felét eltalálták – talán abból indultak ki, hogy önök melyik címre kattintottak volna rá. De telitalálatot aligha értek el.

Vajon miért? Mit hagytak ki a számításból? Az emberi viselkedés mely összetevőjét nem látták át? És milyen tanulságot vonhatnak le a hibákból?

Ilyen kérdéseket rendszerint akkor teszünk fel, ha az előrejelzéseink nem jöttek be.

De nézzük csak, milyen nehéz általános érvényű következtetésekre jutni a *Globe* szalagcímeiből! Az első tesztkérdésben egyetlen szó cseréje („ez a drón” helyett „a SnotBot drón”) hozta meg a nagyarányú győzelmet. Ez azt sugallhatja, hogy a részletgazdagabb cím a nyerő. Csakhogy a második párban veszített a keresőszót konkretizáló „puha tojás” szalagcím, a negyedik párban pedig a „bankot robant” kifejezés verte a számszerű összeget, a 179 ezer dollárt. Ez a két példa azt sugallhatja, hogy a szlengkifejezések a nyerők, csakhogy a harmadik párban alulmaradt a „csajozóverseny”.

Az A/B-tesztelésből nagyrészt azt a tanulságot szűrhetjük le, hogy óvakodjunk az általános érvényű tanulságoktól. Clark Benson, az A/B-tesztelést szalagcímválasztásra és honlaptervezésre is használó szórakoztató és hírportál, a ranker.com vezérigazgatója így nyilatkozott erről: „Vég-eredményben nem vehetünk biztosra semmit. Szó szerint mindent tesztelni kell!”

A tesztelés kitölti az emberi természettel kapcsolatos tudásunk hézagait. Ilyen hézagok mindig is lesznek. Ha élettapasztalatunk megúgná a helyes választ, akkor a tesztelésnek nem lenne semmi haszna. De nem súgja meg, ezért van haszna.

Az A/B-tesztelés jelentőségének másik oka az, hogy lát-szólag apró változtatások is óriási hatást képesek kiváltani. Ismét Bensont idézem: „Újra és újra megdöbbenek azon, hogy milyen óriási értékkel bírnak a tesztelésben mindenféle icipici tényezők.”

A Google 2012 decemberében megváltoztatta hirdetéseit: hozzájuk csapott egy kockában elhelyezett, jobbra mutató nyilat.

Szállodák

www.example.com

Különleges árak hó végéig!

Foglalási díj nincs, foglaljon szobát most!



Dublini szállodák

www.example.com

Böngésszen át több száz dublini szállodát, és válogasson köztük ár, hely és a felhasználók értékelései szerint!



Szállodák Írországban

www.example.com

Hasonlítsa össze több ezer szálloda árait Írország-szerte!



AdChorus

Vegyük észre, milyen bizarr ez a nyíl! Jobboldalt, amerre mutat, az égvilágon semmi sincs. Eleinte sok Google-felhasználó is kritikusan viszonyult hozzá: nem értették, miért kell belerakni a hirdetésekbe ilyen értelmetlen nyílakat.

Nos, a Google féltve őrzi üzleti titkait, így nem mondják meg, hogy pontosan mekkora hasznát veszik a jobbra mutató nyílnek. Azt viszont elárulták, hogy győztesen került ki az A/B-tesztelésből. Vagyis azért csapták hozzá a hirdetésekhez, mert így sokkal többen kattintanak rá azokra. Az értelmetlenné tűnő, apró változtatás egy rakás pénzt hozott a Google és hirdetési partnerei konyhájára.

Hogy fedezhetjük fel ezeket az apró, de hatalmas profitot termelő trükköket? Úgy, hogy tesztelünk egy csomó mindent, még akkor is, ha olykor triviálisnak tűnnek. A Google-felhasználók már sokszor észleltek olyat, hogy a hirdetések egy apró módosítás után visszanyerik korábbi formájukat. Ilyenkor az történt, hogy tudtukon kívül a kísérleti csoport tagjai lettek egy A/B-teszten, ami abban merült ki, hogy látták ezeket az apró változásokat.

Centering Experiment (Didn't Work) =
Középre zárás kísérlet (nem jött be)

Green Star Experiment (Didn't Work) =
Zöld csillagos kísérlet (nem jött be)

New Font Experiment (Didn't Work) =
Új betűtípusos kísérlet (nem jött be)

A fenti három változtatás nem jutott el a tömegekhez. Veszítettek. De így is részesei voltak a kiválasztódás folyamatának. A „kattintható” nyílhoz vezető rögzös út ronda csillagokkal, béna szövegképekkel és hatásvadász betűtípusokkal van kikövezve.

Mókás dolog tud lenni azt tippelgetni, hogy mitől kattintanak az emberek. A Demokrata Párt híveit pedig talán az is jóleső érzéssel tölti el, hogy Obama a tesztelés révén több kampánytámogatáshoz jutott. De az A/B-tesztelésnek is megvan a maga sötét oldala.

Adam Alter *Ellenállhatatlan* című kiváló könyvében arról ír, hogy manapság társadalmunkban elszaporodtak az addiktív viselkedésformák. Sokan vannak, akik egyre nehezebben tudják kikapcsolni az internet egyes tartalmait.

Kedvenc adatforrásom, a Google-keresések halmaza szolgál bizonyos nyomokkal arra vonatkozóan, hogy az emberek mit találnak a leginkább addiktívnek. A Google szerint zömmel továbbra is azok a dolgok okoznak függőséget, amelyekkel már hosszú évtizedek óta küzdünk – például a drog, a szex és az alkohol. De az internet is elkezdte képviseltetni magát a listán: a „pornó” és a „Facebook” immár szerepel a tíz leggyakoribb függőség között, amelyről az emberek beszámolnak.

A GOOGLE-NAK LEGGYAKRABBAN BEVALLOTT FÜGGŐSÉGEK, 2016

Drog	Alkohol	Szerencsejáték
Szex	Cukor	Facebook
Pornó	Szerelem	

És abban, hogy az internet ilyen átkozottul addiktív, az A/B-tesztelés is szerepet játszhat.

Az *Ellenállhatatlan* című könyv idézi Tristan Harris „dizájnethikust”, aki elmagyarázza, miért tudnak olyan nehezen ellenállni bizonyos oldalaknak az emberek: „ezrek állnak lesben a képernyő másik oldalán, akiknek mind az a dolguk, hogy megingassák az önuralmunkat”.*

* Adam Alter: *Ellenállhatatlan. Hogyan okoz függőséget a technológia?* HVG Könyvek, 2018, 11. Pétersz Tamás fordítása.

És ezek az emberek az A/B-tesztelés módszerét alkalmazzák.

Ezzel a módszerrel a Facebook rájöhet például, hogy ha egy bizonyos gombot bizonyos színűre „fest”, azzal eléri, hogy az emberek sűrűbben járjanak vissza az oldalra. Így az a bizonyos gomb olyan színű lesz. Utána rájöhetnek arra is, hogy egy bizonyos betűtípus többször csábítja vissza a felhasználókat az oldalra, így a szöveg attól kezdve olyan betűtípusban fog megjelenni. Rájöhetnek arra is, hogy ha egy bizonyos időpontban e-mailt küldenek az embereknek, azok többször fognak ellátogatni az oldalra. Ezért el is küldik az embereknek azt az e-mailt, abban a bizonyos időpontban.

A Facebook így elég hamar olyan weboldallá válik, amelyet sikerült az emberek Facebookon töltött idejének maximalizálására optimalizálni. Más szóval: ha elég sok A/B-tesztet végzünk, a tesztek győzteseinek jóvoltából a weboldalunk addiktívva válik. Efféle visszacsatolásról a dohánygyárok csak álmodhattak.

Az A/B-tesztelés egyre inkább beépül a számítógépesjáték-ipar eszköztárába is. Adam Alter is kitér rá könyvében, hogy a World of Warcraft A/B-teszteli a játék különféle verzióit. A tervezők küldetésekkel bízzák meg a játékosok különböző csoportjait. Az egyik csoportnak meg kell ölnie valakit. A másik csoportnak meg kell mentenie valakit. Aztán megnézik, hogy melyik csoportból folytatják többet a játékot. Kiderülhet például, hogy a megmentős küldetés után 30 százalékkal gyakrabban térnek vissza a játékosok. Ha sok-sok küldetést letesztenek, a győztesek 30 százaléka összeadódik, és a végeredmény egy olyan játék lesz, amelynek hatására felnőtt férfiak tömegei gubóznak be a szülői ház pincéjében.

Ha ez némileg felkavarja önt, akkor egyetértünk.

A könyv vége felé fogunk is még beszélni ennek és a Big Data más vonatkozásainak etikai vonatkozásairól. De egyvalami biztos: a kísérletezés kulcsfontosságú eleme az adatudósok eszköztárának. És ebben az eszköztárban ott lapul a kísérletezés egy másik fajtája is, amelynek révén már igen sokféle kérdést feltettek. Például azt, hogy tényleg működnek-e a tévéreklámok.

A TERMÉSZET KEGYETLEN, DE TANULSÁGOS KÍSÉRLETEI

2012. január 22-ét írunk. Az amerikaifoci-bajnokság egyik csoportja, az American Football Conference (AFC) döntőjét, amely egyben a Super Bowl elődöntője, a New England Patriots és a Baltimore Ravens játssza.

Egy perc van hátra a mérkőzésből. A baltimore-iak hátrányban vannak, de náluk a labda. A következő hatvan másodpercben eldől, hogy melyik csapat fog játszani a Super Bowlban. A játékosok a következő hatvan másodpercben beírhatják magukat a történelembe. És a meccs utolsó percében történni fog még valami, ami egy közgazdász számára sokkal lényegesebb. Az utolsó hatvan másodpercben ugyanis végre-valahára, egyszer s mindenkorra kiderül, hogy működnek-e a reklámok.

Az az elképzelés, hogy a hirdetés javítja az eladási mutatókat, nyilván kulcsfontosságú tényező a gazdaságunkban. Bizonyítani viszont örületesen nehéz. Olyannyira, hogy szinte tankönyvi például szolgál a korreláció, illetve az ok-sági összefüggés közti különbségtétel problémájára.

Kétségtelen, hogy a legtöbbet reklámozott termékek fognak a legjobban. A Twentieth Century Fox például 150 millió dollárt költött az *Avatar* marketingjére, és a film

minden idők legnagyobb kasszasikere lett. De a 2,7 milliárd dolláros jegybevételből vajon mennyi köszönhető a bivalyerős marketingnek? A Twentieth Century Fox feltehetően részben azért költött ennyit a promócióra, mert tudták, hogy vonzó terméket állítottak elő.

A cégek azt hiszik, hogy tudják, mennyire hatékonyak a hirdetéseik. A közgazdászok viszont szkeptikusak e téren. Steven Levittet, a Chicagói Egyetem közgazdászprofesszorát egyszer tanácsadásra kérte fel egy elektronikai cég. Levitt nem volt elkápráztatva, amikor arról próbálták meggyőzni, hogy tudják, mennyire válnak be a hirdetéseik. Nem értette, hogyan lehetnek ilyen biztosak ebben.

Elmagyarázták neki, hogy az apák napját megelőző napokban mindig sokat költenek tévéreklámra – és lám, minden évben akkor tetőznek az eladási mutatóik. Na jó, talán csak azért, mert olyankor sok gyerek vesz valami elektronikus kutyut az apjának, mindenféle reklámtól függetlenül.

„Ami az oksági összefüggést illeti, teljesen fordítva ültek a lovon” – mondta egy előadásában Levitt. Vagy legalábbis elképzelhető, hogy fordítva ültek a lovon: nem tudjuk, hogy tényleg így volt-e. „Nagyon nehéz probléma ez” – tette hozzá Levitt.

De akármennyire fontos lenne is, hogy ezt a problémát megoldják, a cégek nemigen hajlanak rá, hogy precíz kísérleteket végezzenek. Levitt professzor megpróbálta rávenni azt a bizonyos elektronikai céget, hogy randomizált kontrollált kísérlet útján derítsék ki, mennyire hatékonyak valójában a tévéreklámjaik. Mivel az A/B-tesztelést egyelőre lehetetlen tévén keresztül alkalmazni, ki kellett volna próbálniuk, hogy bizonyos területeken mit tapasztalnak reklám nélkül.

A cég reakciója a következő volt: „Maga megőrült? Nem tehetjük meg, hogy húsz helyen ne hirdessünk! A vezér-

igazgató kinyírna minket!” Így ért véget Steven Levitt és az elektronikai cég együttműködése.

Kanyarodjunk most vissza ahhoz a bizonyos Patriots-Ravens mérkőzéshez! Hogyan segíthet egy amerikaifoci-meccs eredménye abban, hogy megállapítsuk, volt-e oksági hatásuk a hirdetéseknél? Nos, egy konkrét cég konkrét reklámkampányának hatásait nem tudjuk mérni az eredmény alapján. De sok nagyszabású reklámkampány átlagos hatásáról már megtudhatunk ezt-azt.

Egy ilyen meccs valójában nem más, mint titkos reklámkísérlet. A dolog úgy működik, hogy mire a bajnoki csoportdöntőket, azaz a Super Bowl-elődöntőket lejátszzák, a cégek már megvásárolták a reklámidőt, és legyártották a Super Bowl-hirdetéseiket. Vagyis amikor eldöntik, hogy milyen hirdetést adjanak le a Super Bowl alatt, még nem tudják, hogy mely csapatok fogják játszani a döntőt.

Pedig a rájátszás eredményei nagy hatással vannak arra, hogy kik fogják nézni a Super Bowlt. A döntőbe jutó két csapat ugyanis rengeteg nézőt hoz magával. Ha a Bostonhoz kötődő New England Patriots nyer a Baltimore Ravens ellen, akkor Bostonban sokkal többen fogják nézni a Super Bowlt, mint Baltimore-ban, és fordítva.

A cégek szempontjából ez olyan, mintha pénzfeldobással dőlné el, hogy a baltimore-iak, illetve a bostoniak tízezrei látni fogják-e a hirdetéseiket. És a pénzfeldobásra csak akkor kerül sor, amikor ők már kifizették és legyártották a reklámspotjaikat.

Most pedig vissza a meccshez, ahol is Jim Nantz, a CBS kommentátora éppen közli velünk a kísérlet végeredményét!

És már jön is Billy Cundiff, hogy döntetlenre hozza ki a mérkőzést, amely így minden valószínűség szerint hosszabbításban fog eldőlni. Cundiff az utóbbi két évben ti-

zenhat kísérletből tizenhat mezőnygólt rúgott. Harminckét yardról kell betalálnia a döntetlenhez. Cundiff lö... Odanézzenek! Odanézzenek! Mellé! És a Patriots most már kihúzza a hátralevő pár másodpercet, így ők utazhatnak majd Indianapolisba, a negyvenhatodik Super Bowlra.

Két héttel később a negyvenhatodik Super Bowl Bostonban 60,3 százalékos, Baltimore-ban pedig 50,2 százalékos nézettséget ért el. Bostonban hatvanezerrel többen látták a 2012-es Super Bowl-hirdetéseket.

Egy évre rá ugyanez a két csapat találkozott az AFC bajnoki döntőjében. Ezúttal a Baltimore nyert, így a 2013-as Super Bowl-hirdetéseket már Baltimore-ban látták többen.

	A 2012-AS SUPER BOWL NÉZETTSÉGE (A BOSTON JÁTSZOTT)	A 2013-AS SUPER BOWL NÉZETTSÉGE (A BALTIMORE JÁTSZOTT)
Boston	56,7%	48,0%
Baltimore	47,9%	59,6%

E két mérkőzés, valamint a 2004 és 2011 közötti Super Bowlok alapján teszteltük Hal Variannal, a Google vezető közgazdászával és Michael D. Smithszel, a Carnegie Mellon közgazdászával, hogy működnek-e a Super Bowl-hirdetések, és ha igen, mennyire. Konkrétan azt néztük meg, hogy ha lejátszák egy film előzetesét a Super Bowl alatt, megugrik-e az eladott mozijegyek száma azokban a városokban, ahol magas volt a bajnoki döntő nézettsége.

Igen, megugrik. Azon városok lakói, amelyeknek csapata kvalifikálja magát a Super Bowlra, lényegesen nagyobb arányban ülnek be a meccs alatt reklámozott filmekre, mint az olyan városokban élők, amelyeknek a csapata épphogy lemaradt a döntőről. A döntős csapatok városaiban többen

látták a reklámot. A döntős csapatok városaiban többen döntöttek úgy, hogy megnézik azt a bizonyos filmet.

Alternatív magyarázatként az is felmerülhet, hogy amennyiben a csapatunk bejut a bajnokság döntőjébe, akkor szívesebben megyünk moziba. Mindazonáltal leteszteltünk néhány hasonló költségvetésű filmet, amelyek nagyjából ugyanakkor kerültek a mozikba, de nem reklámozták őket a Super Bowl alatt. Ezeket a filmeket nem nézték meg többen a döntős csapatok városaiban.

Oké, amint azt önök is sejthették, a reklámok működnek. Ez nem túl meglepő.

De nem csupán arról van szó, hogy működnek. A reklámok hihetetlenül hatékonyak. Olyannyira, hogy amikor megláttuk az eredményeket, újra ellenőriztük, aztán másodszor, aztán harmadszor is ellenőriztünk mindent, hogy biztosan helyesek-e a számok – annyira hatalmas volt a megtérülés. A mintánkban szereplő filmek Super Bowl-reklámspotjai átlagosan mintegy 3 millió dollárba kerültek. A megnövekedett jegyeladások pedig 8,3 milliót hoztak, vagyis a befektetett összeg 2,8-szorosát.

Az eredményt két másik közgazdász, Wesley R. Hartmann és Daniel Klapper is megerősítette, akiknek tőlünk függetlenül és még előttünk támadt hasonló ötletük. Ők a Super Bowl alatt sugárzott sör- és üdítőreklámokat vizsgálták, és szintén figyelembe vették azt, hogy a két döntős csapat városaiban több emberhez jut el a reklám. Ők 2,5-szörös megtérülést mutattak ki. Az ő tanulmányuk és a miénk ugyanarra jutott: bármilyen drágák ezek a Super Bowl-reklámok, olyan hatékonyan pörgetik fel a keresletet, hogy a cégek voltaképpen jóval a valós értékük alatt költenek rájuk.

És vajon mit szűrhetnek le mindebből az elektronikai cégnél dolgozó barátaink, akik úgy megütköztek Levitt

professzor tanácsán? Talán azt, hogy a Super Bowl-reklám költséghatékonyabb, mint bármely más hirdetési forma. És talán azt is, hogy alighanem mégiscsak jó ötlet az apák napja előtti reklámkampány.

A Super Bowl-kísérlet egyik előnye az volt, hogy nem kellett kísérleti csoportra és kontrollcsoportra osztani az embereket. A futball-labda véletlenszerű pattogása döntötte el, hogy melyik alany hova kerüljön. Vagyis az egész természetes úton történt. Hogy ez miért előny? Hát azért, mert a nem természetes körülmények között zajló, random kontrollált kísérletek szuperhatékonyak ugyan, és könnyebb is elvégezni őket a mi digitális korunkban, de még ma sem mindig van rá módunk.

Előfordul, hogy nem kapjuk össze magunkat időben. Előfordul – mint az elektronikai cég esetében, amely nem akarta kísérleti úton tesztelni a hirdetési kampányát –, hogy már túl sokat fektettünk a dologba, ezért nem akarjuk próbára tenni, hogy tényleg jó-e.

Előfordul, hogy egyszerűen nincs mód a kísérletezésre. Tegyük fel, hogy az érdekel minket, miként reagál egy ország a vezetője elvesztésére. Háborút indít? Összeomlik a gazdaság? Vagy nem nagyon változik semmi? Nyilvánvalóan nem áll módunkban, hogy megfelelő számú államfőt és miniszterelnököt legyilkoljunk, aztán megnézzük, hogy mi történik. Ez nem csupán lehetetlen, hanem morálisan is kifogásolható vállalkozás lenne. Az egyetemek az évtizedek során létrehozták a maguk független etikai bizottságait, vagyis „intézményi felülvizsgálati bizottságait” (IRB), hogy ezek döntsék el, etikus-e egy tervezett kísérlet.

Mit tehetünk tehát, ha kíváncsiak vagyunk egy bizonyos forgatókönyv oksági hatásaira, de a kísérlet elvégzése etikátlan lenne, esetleg más okból megvalósíthatatlan? Nos,

ilyenkor jól jöhet az, amit a közgazdászok – akik annyira tágan értelmezik a „természetesség” kritériumát, hogy abba még egy amerikaifutball-meccs is belefér –, úgy hívnak, hogy természetes kísérlet.

Akár tetszik, akár nem (na jó, nyilván nem), az élet igen jelentős részben véletlenszerű. Senki sem tudja biztosan, hogy micsoda vagy kicsoda irányítja az univerzumot. Ám egy dolog egészen biztos: akárki(k) rendezi(k) a műsort – legyen szó akár a kvantummechanika törvényeiről, akár Istenről, akár egy pattanásos, alsógatyás kölyökről, aki a számítógép előtt ülve szimulálja a világegyetemet –, nem kell jóváhagyatnia vagy jóváhagyatniuk a dolgot az IRB-vel.

A természet szüntelenül kísérletezik rajtunk. Meglőnek két embert. Az egyik golyó megáll egy létfontosságú szerv előtt. A másik nem. Az ilyen balszerencsék miatt nem fair az élet. Ám ha ez bárkit is vigasztal, a közgazdászok életét valamelyest megkönnyíti az efféle balszerencse. A közgazdászok arra használják fel az élet önkényességét, hogy oksági kapcsolatokat teszteljenek általuk.

Az eddigi 43 amerikai elnök közül 16-an lettek komoly merényletkísérlet áldozatai, és négyüket meg is gyilkolták. Hogy miért maradt életben, aki életben maradt, az nagyrészt random okokra vezethető vissza.

Vessük össze például John F. Kennedy és Ronald Reagan esetét! Mindkettőjüknek az egyik legérzékenyebb szerve felé tartott a golyó. JFK agyát szét is loccsantotta, és az elnök nem sokkal később meghalt. A Reagan testébe hatoló lövedék viszont pár centivel a szíve előtt megállt, így az orvosok meg tudták menteni az életét. Reagan tehát életben maradt, JFK pedig meghalt. Itt aztán tényleg nem volt „se lába, se keze a »miértnek-azértnek«”.*

* William Shakespeare: *Tévedések vígjátéka*. Szász Imre fordása. (A ford.)

Politikai vezetők elleni merényletek világszerte előfordulnak, és mindenütt a vakszerencsén múlik, hogy életben maradnak-e. Vessük csak össze a csecsen Ahmad Kadirov és a német Adolf Hitler esetét! Mindketten centiméterekre voltak egy teljes mértékben működőképes bombától, de csak Kadirov halt meg. Hitler ugyanis módosította a programját: a tervezettnél néhány perccel hamarabb hagyta el az aláaknázott helyiséget, hogy elérjen egy vonatot, így aztán megúszta.

Nekünk pedig kapóra jön a természet rideg véletlenszerűsége – Kennedy meghal, Reagan pedig nem –, hogy megnézhessük, mi történik általában egy ország vezetőjének meggyilkolása után. Két közgazdász, név szerint Benjamin F. Jones és Benjamin A. Olken, pontosan ezt vizsgálta. A kontrollcsoport itt bármely ország egy épphogy megúszott merénylet után – például az Egyesült Államok a nyolcvanas évek közepén. A kísérleti csoport pedig bármely ország egy sikeres merénylet után – például az Egyesült Államok a hatvanas évek közepén.

Milyen hatást vált ki tehát, ha az ország első emberét meggyilkolják? Jones és Olken arra jutott, hogy a sikeres merényletek drámaian megváltoztatják a világtörténelmet: egészen más pályára állítanak egy-egy országot. Az új vezetők az addig békében élő országokat háborúba viszik, az addig háborúskodó országok számára pedig elérik a békét. Az új vezetők az addig prosperáló országokat a csőd szélére juttatják, az addig a csőd szélén tántorgó országokat pedig felvirágoztatják gazdaságilag.

Ami azt illeti, a természetes kísérlet, amely az államfők elleni merényleteket vizsgálta, romba döntötte a megelőző évtizedek közkeletű vélekedését arra vonatkozóan, hogy miként működnek az országok. A közgazdászok addig erősen hajlottak arra, hogy az államfők csak impotens,

magatehetetlenül sodródó bábok, akiket külső erők kedvük szerint lökdösnek ide-oda. Jones és Olken elemzése szerint viszont egyáltalán nem ez a helyzet.

A merényletkísérletek e vizsgálatát sokan nem tekintnék jó példának a Big Data alkalmazására. A meggyilkolt vagy majdnem meggyilkolt vezetők száma kétségkívül csekély, ahogy a kirobbant vagy ki nem robbant háborúk száma is. A gazdaság pályájának leírásához szükséges gazdasági adathalmazok pedig nagyok ugyan, de többnyire a digitalizáció előtti korból valók.

Az efféle természetes kísérletek – bár egyelőre szinte kizárólag a közgazdászok alkalmazzák a módszert – mindezzel együtt is igen hatékonyak, és egyre fontosabbá válnak majd az egyre számosabb, egyre jobb, egyre nagyobb adathalmazok korában. Olyan eszköz ez, amelyről az adattudósok nem mondhatnak le tartósan.

És igen, amint az mostanra alighanem világossá vált, a közgazdászok kulcsfontosságú szerepet játszanak az adattudomány fejlődésében. Legalábbis szeretném hinni, hogy így van, hiszen én is közgazdásznak tanultam.

Vajon hol bukkanhatunk még természetes kísérletekre, azaz olyan helyzetekre, amikor véletlenszerű események kísérleti csoportra és kontrollcsoportra osztják az embereket?

A legegyszerűbb példa a lottó, amelyet épp ezért imádnak a közgazdászok – mármint nem lottózni, mert azt irracionális dolognak tartják, hanem vizsgálni. A dolog, ugyebár, úgy működik, hogy ha a hármas számot viselő pingponglabda kerül legfelülre, akkor Mr. Jones gazdagszik meg, ha pedig a hatos, akkor Mr. Johnson.

A talált pénz oksági hatásait vizsgáló közgazdászok a lottónyerteseket hasonlítják össze azokkal, akik vesznek ugyan lottószelvényt, de nem nyernek. Ezek a tanulmá-

nyok rendszerint azt mutatják ki, hogy a lottónyeremény rövid távon nem teszi boldoggá az embert, hosszú távon azonban igen.*

A közgazdászoknak a lottósorsolás véletlenszerűsége akkor is jól jöhet, ha azt akarják vizsgálni, hogy miként változik meg az ember élete, ha a szomszédja meggazdagodik. Az adatok tanúsága szerint a szomszéd lottónyereményének lehet hatása az életünkre. Példának okáért, ha a szomszédunk megnyeri a főnyereményt, igen valószínű, hogy vásárolni fogunk egy drága kocsit, például egy BMW-t. Hogy miért? A közgazdászok szerint szinte biztosan azért, mert irigyeljük a szomszédot, aki szintén vett egy drága kocsit. Ezt bizvást írhatjuk az emberi természet számlájára: ha Mr. Johnson meglátja, hogy Mr. Jones vadonatúj BMW-vel furikázik, akkor neki is kell egy.

De sajnós Mr. Johnsonnak sokszor nem telik új BMW-re. Ezért tapasztalták a közgazdászok azt, hogy a lottónyertesek szomszédai jóval nagyobb eséllyel mennek csődbe, mint mások. Az ő esetükben egyszerűen lehetetlen lépést tartani Jonesékkal.

A természetes kísérleteknek ugyanakkor nem kell feltétlenül annyira elrendezett módon véletlenszerűnek lenniük, mint a lottósorsolásnak. Ha egyszer elkezdjük keresni a véletlenszerűséget, mindenütt meg fogjuk találni – és a segítségével megérthetjük, hogyan is működik a világunk.

Az orvosok is természetes kísérlet alanyai. A kormány ugyanis időnként, lényegében önkényes okokból, megváltoztatja az olyan páciensek utáni térítés kiszámításának módját, akiket az állami egészségbiztosítási program, a Medicare terhére kezelnek. Ilyenkor az orvosok azzal

* Egy híres, 1978-as tanulmány állítása, amely szerint a lottónyeremény nem boldogít, nagyjából megdől.

szembesülnek, hogy egyes megyékben bizonyos eljárásokért többet kapnak, mint addig, más megyékben pedig kevesebbet.

Két közgazdász, név szerint Jeffrey Clemens, valamint egykori osztálytársam, Joshua Gottlieb tesztelte az önkényes tarifamódosítás hatásait. Azt vizsgálták, hogy mindig ugyanolyan ellátásban részesül-e a beteg, azaz mindig a leginkább szükségesnek ítélt terápiát alkalmazzák-e az orvosok, vagy esetleg hat rájuk az anyagi ösztönzés.

Az adatok egyértelműen azt mutatják, hogy az orvosokat motiválhatják az anyagiak. Az olyan megyékben ugyanis, ahol magasabb a térítés összege, egyes orvosok lényegesen többször alkalmazzák a jobban finanszírozott eljárásokat: például több a szürkehályogműtét, a vastagbél-tükrözés és az MRI-vizsgálat.

És most jön a nagy kérdés: jobban járnak-e a páciensek a drágább kezelésekkal? Clemens és Gottlieb csak „csekély mértékű egészségügyi hatásokról” számolt be, és azt sem tapasztalták, hogy a halálozási arány statisztikailag szignifikáns mértékben változott volna. A természetes kísérlet eredménye tehát azt sugallja, hogy ha az orvosokat anyagiilag ösztönözzük bizonyos eljárások elvégzésére, akkor egy részük többször fog olyan terápiát alkalmazni, amely nem javítja érdemben a páciens egészségi állapotát, és nem hosszabbítja meg az életét.

A természetes kísérletek valódi élet-halál kérdésekre is választ adhatnak. És olyanokra is, amelyeket a fiatalok egy része annak érez.

A Stuyvesant Középiskolának (közismert becenevén „Stuy”-nak) egy 150 millió dollárt érő, tízemeletes, rozsdabarna, a Hudson-folyóra néző téglalapú épület ad otthont Alsó-Manhattanben, néhány saroknyira a World Trade Centertől. Ha

egyetlen szóval akarjuk leírni a Stuyt, az a lenyűgöző. A diákok ötvenötféle egyetemi szintű órára járhatnak, hétféle nyelvet tanulhatnak, és a választható tárgyak listáján a zsidó történelem, a science fiction, valamint az ázsiai-amerikai irodalom is szerepel.

Az itt végzetek mintegy negyedét veszik fel a Borostyán Liga* valamelyik egyetemére vagy más, hasonlóan patinás felsőoktatási intézménybe. A Stuyvesant Középiskolában tanult Lisa Randall, a Harvard fizikaprofesszora, David Axelrod, Obama egyik fő politikai tanácsadója, Tim Robbins Oscar-díjas színész és Gary Shteyngart regényíró. Tanévzáró beszédet pedig mondott már a végzősöknek Bill Clinton, Kofi Annan és Conan O'Brien is.

A Stuyvesant tanrendjénél és az itt végzetek névsoránál talán csak a tandíj összege elképesztőbb: nulla, azaz zéró dollár! Ez ugyanis önkormányzati iskola, méghozzá alighanem a legjobb az országban. Sőt, nemrég készült egy felmérés, amely 300 ezer diák és szülő összesen 27 millió értékelése alapján rangsorolta az ország összes önkormányzati fenntartású gimnáziumát – a Stuy az első helyen végzett! Nem csoda hát, hogy az ambiciózus, középosztálybeli New York-i szülők és az ő nem kevésbé ambiciózus utódaik odavannak a Stuy brandjéért.

Ahmed Yilmaz** szemében, aki egy biztosítási ügynök és egy tanárnő fiaként nőtt fel Queensben, a Stuy volt „a középiskola”. Az ő szavait idézem: „A munkásosztálybeli, illetve bevándorló családok kiugrási lehetőségnek tekintik a Stuyt. Ha a gyereked oda jár, nyugodt lehetsz, mert be fog kerülni egy rendes egyetemre, amelyik benne van a legjobb húszban. A család gondjai pedig megoldódnak.”

* Az Egyesült Államok nyolc legszínvonalasabb elit egyeteme. (A szerk.)

** A nevét és néhány adatot megváltoztattam.

Akkor hát hogyan lehet bekerülni a Stuyvesant Középiskolába? Nos, az kell hozzá, hogy New York City öt kerületének valamelyikében lakj, és elérj a felvételin egy bizonyos pontszámot. Ennyi. Semmi ajánlólevél, semmi esszé, semmi apucitól örökölt hely, semmi pozitív diszkrimináció. Egy nap, egy teszt, egy pontszám. És ha a pontszámod eléri a küszöböt, bent vagy.

Minden év novemberében úgy 27 ezer ifjú New York-i felvételizik az iskolába. A túljelentkezés brutális. A tesztet megírók alig 5 százaléka kerül be.

Ahmed Yilmaz elmondta, hogy az édesanyja „szétgürizte az agyát”, és a kevéske pénzét mind az ő felkészítésébe ölte. Az ifjú Ahmed hosszú hónapokon át minden hétköznap délutánt és a teljes hétvégét is tanulással töltötte. Biztosra vette, hogy bekerül, és máig emlékszik a napra, amikor megkapta az eredményt tartalmazó borítékot. Két pont hiányzott.

Megkérdeztem tőle, milyen érzés volt. Így felelt: „Szerinted milyen érzés, hogy darabokra hullik a világ, pedig még csak elemista kölyök vagy?”

Pedig a vigaszdíja minden volt, csak ócska nem – a Bronx Science, amely szintén igen magasra rangsorolt, exkluzív önkormányzati középiskola. De nem a Stuy. Yilmaz pedig úgy érezte, hogy amolyan szakiskolába került, amely műszaki érdeklődésűeknek való. Négy évvel később nem vették fel a Princetonra. Beiratkozott hát a Tuftsra, és utóbb többször is pályát módosított. Ma egy technológiai cég meglehetősen sikeres alkalmazottja, de szerinte „agyzsibasztó” a munkája, és nem fizetik meg úgy, ahogy szeretné.

Yilmaz most, több mint egy évtized elteltével, bevallja: néha eszébe jut, hogy vajon miként alakult volna az élete, ha a Stuyba járhat. „Minden más lenne” – mondja. „Szó szerint egészen más lenne mindenki, akit csak ismerek.”

Úgy véli, hogy ha a Stuyvesant Középiskolában végez, talán több pontot ér el a SAT-teszteken, azaz a központi egyetemi felvételi teszten, bejut a Princetonra vagy a Harvardra (mindkettőt lényegesen jobbnak tartja a Tufts-nál), és talán sikerült volna jövedelmezőbb vagy a szívéhez közelebb álló munkát találnia.

Ha az ember azon járhatja az agyát, hogy „Mi lett volna, ha?, az igen szórakoztató tud lenni, de akár önkínzással is felérhet. Milyen lenne az életem, ha ráhajtok arra a lányra vagy arra a fiúra? Ha elvállalom azt a munkát? Ha abba az iskolába járhattam volna? De az ilyen „Mi lett volna, ha?”-kérdésekre nincs válasz. Az élet nem videojáték. Nem játszhatjuk újra, mindenféle alternatív forgatókönyvek szerint, amíg meg nem kapjuk a kívánt eredményt.

Idekívánczik egy magvas idézet Milan Kunderától, aki *A lét elviselhetetlen könnyűsége* című regényében így ír: „Az ember sohasem tudhatja, mit akarjon, mert csak egy élete van, s azt semmiképp nem tudja összehasonlítani az előző életeivel, vagy megjavítani az elkövetkező életei során. (...) Nincs lehetőségünk ellenőrizni, melyik döntésünk jobb, mert összehasonlításra nincs módunk.”*

Yilmaznak sem lesz másik élete, amelyben valahogy sikerül összekaparnia azt a bizonyos két pontot azon a bizonyos teszten. Mi viszont talán bepillantást nyerhetünk abba, hogy mennyiben lehetne vagy nem lehetne más az élete, ha a Stuyvesant Középiskola jó sok egykori diákját megvizsgáljuk.

A buta, naiv módszer az lenne, hogy az összes olyan embert, aki a Stuyvesantba járt, összehasonlítjuk az ös-

* Milan Kundera: *A lét elviselhetetlen könnyűsége*. Európa Kiadó, 1994, 15. Körtvélyessy Klára fordítása. (A szerk.)

szes olyan emberrel, aki nem járt oda. Kielemezhetnénk az AP-teszteken, vagyis az egyetemi szintű középiskolai teszteken, illetve a SAT-teszteken nyújtott teljesítményüket, majd azt, hogy hová vették fel őket. Ha mindezt megtennénk, kiderülne, hogy a Stuy tanulói sokkal több pontot érnek el a központi felvételikén, és lényegesen jobb egyetemekre veszik fel őket. De amint arról már meggyőződhattunk ebben a fejezetben, az ilyen típusú bizonyíték önmagában nem meggyőző. Lehet, hogy azért teljesítenek ennyivel jobban, mert a Stuy-eleve sokkal jobb képességű diákokat vonz. A korreláció ebben az esetben nem bizonyítja az oksági összefüggést.

Ahhoz, hogy teszteljük a Stuyvesant Középiskola *oksági* hatásait, azokat a csoportokat kell összevetnünk, amelyek majdnem azonosak, csak egyetlenegy dologban különböznek: az egyik átélte a „Stuy-kísérletet”, a másik pedig nem. El kell végeznünk egy természetes kísérletet. De hogyan?

A megoldás: az olyan diákok, mint Ahmed Yilmaz, akik nagyon-nagyon közel jártak a felvételi küszöbhez.* Azok alkotják a kontrollcsoportot, akik épphogy a „vonal” alatt maradtak, a kísérleti csoportot pedig azok, akik épphogy elérték a küszöböt.

Nemigen van okunk arra gyanakodni, hogy a vonal két oldalán lévők tehetsége, illetve motiváltsága között jelentős különbség van. Végso soron miért ér el valaki egy-két

* Amikor olyan egykori felvételizőket kerestem, akik nagyon közel voltak a határhoz, igencsak megdöbbsenttem, hogy milyen sokan vannak (a huszonevesektől egészen az ötvenes korosztályig) azok, akik mindmáig jól emlékeznek erre a kamaszkori élményre, és úgy beszélnek róla, mint valami nagy drámáról. Közéjük tartozik Anthony Weiner egykori kongresszusi képviselő és New York-i főpolgármester-jelölt is, aki elmesélte, hogy egyetlen ponttal maradt le a Stuyvesantról. „Nem kellett nekik!” – mondta a telefoninterjúban.

ponttal többet egy teszten, mint a másik? Talán az, aki kevesebb pontot gyűjtött, tíz perccel kevesebbet aludt, vagy kevésbé volt kiadós a reggelije. Talán annak, aki több pontot gyűjtött, eszébe jutott egy különösen nehéz szó, mert három éve hallotta a nagymamájától beszélgetés közben.

A természetes kísérleteknek ez a kategóriája – amelyekben egyértelműen meghúzott, számszerűsíthető küszöbvel dolgozunk – olyan jó hatásfokkal működik, hogy a közgazdászok nevet is adtak neki: regressziós diszkontinuitásnak hívják. Ha van egy pontos számérték, amely két csoportra osztja az embereket (tehát megszakítja a folytonosságot, vagyis diszkontinuitást idéz elő), akkor össze lehet hasonlítani a küszöbhez közel állók későbbi sorsát. Ez a regresszió, azaz a véletlen változók közti kapcsolat elemzése.

Két közgazdász, M. Keith Chen és Jesse Shapiro, azt próbálta felmérni, hogy milyen hatással vannak a durva fogva tartási körülmények a jövőbeni bűnelkövetésre. Nekik az jött kapóra, hogy a szövetségi börtönökben van egy éles választóvonal: az őrizetese a bűncselekmény jellege, illetve a korábbi priuszuk alapján kapnak egy pontszámot, és ez határozza meg a fogva tartásuk körülményeit. Akik elérik a küszöböt, azok magas biztonsági fokozatú fegyintézetbe kerülnek, azaz kevesebb kapcsolatuk lesz másokkal, kisebb lesz a mozgásszabadságuk, és valószínűsíthető, hogy több erőszakot szenvednek el az őröktől, illetve a társaiktól.

Itt sem lenne fair, ha a magas biztonsági fokozatú börtönök teljes népességét vetnénk össze az alacsony biztonsági fokozatú börtönök teljes népességével. A magas biztonsági fokozatú fegyintézetekben nyilván több a gyilkos és az erőszaktevő, az alacsony biztonsági fokozatúakba pedig többen kerülnek kábítószerrel való visszaélés vagy éppen piti lopás miatt.

Azoknak viszont, akik közvetlenül az éles választóvonal felett, illetve alatt helyezkednek el, lényegében azonos a priuszuk és a háttérük. Amiatt a nyomorult egy pont miatt mégis egészen másféle börtöntapasztalat jut nekik.

Az eredmény? Chen és Shapiro megállapította, hogy ha valaki durvább körülmények között raboskodik, nagyobb a valószínűsége annak, hogy a szabadulása után újabb bűncselekményt követ el. Vagyis a kemény börtönélet nem riasztja el az őrizetéseket a bűnöző életmódtól: mire visszajutnak a kinti világba, inkább megedződnek, és erőszakosabbá válnak.

De vajon mit mutatott a regressziós diszkontinuitás a Stuyvesant Középiskola esetében? A tanulmányt az MIT és a Duke közgazdászokból álló kutatócsoportja végezte, név szerint Atila Abdulkadiroglu, Joshua Angrist és Parag Pathak. Összehasonlították a vonal két oldalán lévő New York-i diákokat, vagyis megnézték több száz olyan felvételizőt, mint Ahmed Yilmaz, aki egy-két ponttal maradt el a küszöbtől, és *nem került be* a Stuyvesantba, valamint több száz olyat, akinek aznap jobb napja volt, és egy-két ponttal a küszöb *feletti* eredményt ért el. A diákok későbbi sikerességét az AP-tesztjeik és a SAT-tesztjeik eredménye alapján, valamint az egyetemük rangsorolása alapján mérték.

Hogy milyen döbbenetes eredményre jutottak, azt egyértelműen jelzi, hogy tanulmányuknak az „Elite Illusion”, vagyis az „Elitillúzió” címet adták. Mekkora hatása volt a Stuyvesant Középiskolának? Semekkora! Nulla, zéró! Spanyolul *nada*, jiddisül *bupkus!* A két csoport AP- és SAT-pontszámai között nem volt érzékelhető különbség, ahogy a nagy tekintélyű egyetemeik között sem.

A kutatók ebből azt szűrték le, hogy a Stuy diákjai kizárólag azért viszik többre az életben, mert a Stuyvesant-

ba eleve jobb képességű gyerekek járnak. Az ember nem a Stuy miatt teljesít jobban az AP-teszten és a SAT-teszten vagy köt ki jobb egyetemen.

„Úgy tűnik, hogy a diákok egy jelentős csoportja esetében a bekerülésért folyó intenzív versengés nem igazolható a megszerzett tudás jobb minőségével” – írják az MIT és a Duke közgazdászai.

Vajon miért nem számít, hogy milyen iskolába jár az ember? Erre további történetek adhatnak választ. Nézzük meg például két New York-i fiatal, Sarah Kaufmann és Jessica Eng esetét, akik már kiskorukban arról álmodoztak, hogy a Stuyba fognak járni. Sarah Kaufmann épphogy elérte a küszöböt: egyetlen ponton múlt. „Nem hiszem, hogy valaha bármi is annyira izgalomba tudna még hozni, mint akkor az, hogy felvettek” – emlékszik vissza. Jessica Eng viszont egy ponttal elmaradt a küszöbtől. Sarah tehát bekerült álmai iskolájába, Jessica viszont nem.

Vajon hogyan alakult ezek után az életük? Mindketten sikeresek a szakmájukban, és szeretik is azt – ami többnyire elmondható azokról a New York-iakról, akik a teszteredmények alapján a legfelső öt százalékba tartoznak. A sors ironiája, hogy Jessica jobban élvezte a középiskolás éveit. A Bronx Science-be járt, amely az egyetlen olyan középiskola, ahol Holokauszt Múzeum is van. Jessica rájött, hogy nagyon szeretne a kulturális emlékek gondozásával foglalkozni, és végül antropológiát hallgatott a Cornellén.

Sarah Kaufmann kissé elveszettnek érezte magát a Stuyban. A diáktársai nagyon rágörcsöltek a jegyeikre, ő pedig úgy érezte, hogy az iskola túl nagy súlyt fektet a tesztekre, ami a tanítás rovására megy. Elmondása szerint „határozottan vegyesek” voltak a tapasztalatai, de tanult belőlük. Rájött például, hogy bölcsészkarra akar menni, mert ott nagyobb súlyt helyeznek a tanításra. Fel is vették álmai

egyetemére, a Wesleyan Universityre. Ott pedig arra jött rá, hogy nagyon szeretne segíteni másokon, úgyhogy jelenleg hátrányos helyzetűek jogsegélyezésére szakosodott ügyvédként dolgozik.

Az emberek idomulnak a tapasztalataikhoz, és akik sikeresek lesznek, azok minden helyzetet a maguk javára tudnak fordítani. Két tényező teszi sikeressé az embert: a tehetsége és a motiváltsága. Tehát nem az, hogy ki mondta a suliban az évváró beszédet, és nem is a többi előny, amelyet a jó nevű iskolák kínálni tudnak.

A Stuyvesant esete csak egyetlen tanulmány, amelynek eredményét valószínűleg gyengíti a tény, hogy a Stuyból épphogy kimaradtak zöme végül szintén jó iskolában kötött ki. Ám egyre több a bizonyíték arra, hogy miközben a jó iskola valóban fontos, abból már vajmi keveset tud profitálni az ember, hogy a lehető legnagyobb iskolába jár.

Vegyük például az egyetemeket. Számít-e, hogy valaki a világ egyik legjobb egyetemén tanul-e (például a Harvardon), vagy „csak” egy jó, megbízható színvonalú tanintézményben (például a Pennsylvániai Állami Egyetemen, azaz a Penn State-en)?

Az iskola rangsorban elfoglalt helyezése ebben az esetben is egyértelműen korrelál azzal, hogy később mennyit fognak keresni az ott végzettek. Az átlagos Harvard-diplomás például évi 123 ezer dollárt keres pályafutása tizedik évében, a Penn State átlagos öregdiákja pedig 87 800 dollárt.

De a korreláció még nem jelent oksági oksági viszont.

Két közgazdász, Stacy Dale és Alan B. Krueger, ötletes módszert dolgozott ki annak tesztelésére, van-e oksági hatásuk az elit egyetemeknek az ott végzettek későbbi pénzkeresési potenciáljára. Dale és Krueger óriási adathalmazmal dolgozott, így nyomon tudták követni a középiskolásokra vonatkozó információk tömkelegét, mint például

hogy ki milyen egyetemre jelentkezett, hová vették fel, hová járt, milyen a családi háttere vagy mennyi a felnőttkori jövedelme.

Hogy legyen kísérleti csoport és kontrollcsoport, Dale és Krueger olyan, hasonló családi háttérű diákokat hasonlított össze, akiket felvettek ugyanabba az intézménybe, de végül mégis két különböző egyetemet választottak. Voltak például olyanok, akik bejutottak a Harvardra, de aztán a Penn State-re jártak – talán azért, hogy közelebb legyenek a barátjukhoz vagy a barátnőjükhöz, talán azért, mert járni akartak egy bizonyos professzor óráira. Vagyis a felvételi bizottság ugyanolyan tehetségesnek találta őket, mint azokat, akik a Harvardra mentek, mégis másféle oktatásban részesültek.

Mi történt tehát, amikor két, hasonló háttérű diákot felvettek a Harvardra, de egyikük a Penn State-et választotta? A kutatók itt is ugyanolyan meglepetésteremtő eredményre jutottak, mint a Stuyvesant esetében: a két ember jövedelme nagyjából megegyezik. Ha tehát a jövőbeni fizetés a mérce, akkor a hasonlóan nagy presztízsű egyetemre felvett, de másikat választó diák nagyjából ugyanoda jut el, mint az, aki az elit egyetem hallgatója lesz.

Újságjaink tele vannak a Borostyán Liga egyetemlein végzett, majd sikert sikerre halmozó emberekről szóló cikkekkel: olyanokról, mint például Bill Gates, a Microsoft alapítója, vagy Mark Zuckerberg és Dustin Moskovitz, a Facebook alapítói, akik mindhárman a Harvardra jártak. (Igaz, ki is maradtak onnan, ami további kérdéseket vet fel a borostyánligás tanulmányok értékére vonatkozóan.)

Ismerünk történeteket olyanokról is, akik elég tehetségesnek bizonyultak ahhoz, hogy felvegyék őket egy borostyánligás egyetemre, de kevésbé patinás tanintézményt választottak, és aztán mégis rendkívül sikeres életet éltek.

Ilyen például Warren Buffett, aki a Wharton Schoolban (a Pennsylvaniai Egyetem üzleti tudományok karán) kezdte meg a tanulmányait, de aztán átment a Nebraskai Egyetemre, mert az olcsóbb volt, továbbá utálta Philadelphiát, és unta a Wharton Schoolt. Az adatok, legalábbis a jövedelmi adatok, azt sugallják, hogy jó döntés lehet, ha az ember kevésbé patinás egyetemet választ, legalábbis Buffettnél és másoknál biztosan bevált.

Azt a címet adtam a könyvnek, hogy *Mindenki hazudik*. Ezalatt elsősorban azt értettem, hogy az emberek azért hazudnak – barátoknak, közvélemény-kutatóknak és önmaguknak –, hogy jobb színben tűnhessenek fel.

De a világ is hazudik nekünk, amikor selejtes, félrevezető adatokat tár elénk. A világ rengeteg sikeres Harvard-diplomást mutat nekünk, sikeres Penn State-diplomából viszont csak kevesebbet, amiből mi azt szűrjük le, hogy hatalmas előnyt jelent, ha valaki a Harvardra jár.

Ha okosan értelmezzük a természet kísérleteit, helyesen értelmezhetjük a világ adatait is – hogy megtudjuk, melyik válhat hasznunkra, és melyik nem.

A természetes kísérletek témája az előző fejezethez is kapcsolódik. Sokszor rá kell közelítenünk a kísérleti csoportra, illetve a kontrollcsoportra: a Super Bowl-kísérletben a városokra, a Medicare-kísérletben a megyékre, a Stuyvesant-kísérletben a küszöbhez közeli pontszámot elért diákokra. Az efféle „zoomoláshoz”, amint arról az előző fejezetben már esett szó, sokszor hatalmas méretű, mindenre kiterjedő adathalmazokra van szükség – amelyek a mai, digitalizálódó világban egyre inkább rendelkezésre állnak. Mivel nem tudhatjuk, mikor szottyan kedve a természetnek egy kis kísérletezéshez, nem tudunk nyélbe ütni egy

kis felmérést az eredmények vizsgálatára. Rengeteg készen kapott adatra van szükségünk ahhoz, hogy tanulhassunk valamit a természetnek ezekből a beavatkozásaiból. Big Datára van szükségünk.

Van még egy fontos dolog, amit el kell mondanom az e fejezetben ismertetett – akár saját, akár természetes – kísérletekről. Ennek a könyvnek a nagy része a világ megértésére összpontosított: olyasmikre, hogy mennyibe került Obamának a rasszizmus, hány férfi meleg valójában, és mennyire bizonytalanok a férfiak és a nők a testükkel kapcsolatban. Az e fejezetben tárgyalt kontrollált, illetve természetes kísérletek viszont praktikusabb indíttatásúak. Az a céljuk, hogy fejlesszék a döntéshozatali képességünket, illetve hogy megtanuljuk, mely beavatkozások működnek, és melyek nem.

A cégek megtudhatják, hogyan tegyenek szert több vásárlóra. A kormányzat megtudhatja, miként ösztönözheti a legjobban az orvosokat. A diákok megtudhatják, hogy mely iskolákban szerzett tudás bizonyulhat később a legértékesebbnek. Ezek a kísérletek a Big Datában rejlő lehetőségeket demonstrálják: azt, hogy a tippelgetés, a közkeletű vélekedések és az ócska kis korrelációk helyébe lépget, és elénk tárhatja, hogy mi az, ami valóban működik – *okási tényezőként*.

III. RÉSZ

BIG DATA: VIGYÁZAT, TÖRÉKENY!

Nagy adat, nagy gond? Ami nem megy

„Seth, Lawrence Summers szeretne találkozni veled.” Ez állt a kissé sejtelmes e-mailben. Lawrence Katztól az egyik PhD-konzulensemtől érkezett. Katz nem tért ki arra, hogy miért érdeklődik Summers a munkám iránt, de később kiderült, hogy nagyon is jól tudta.

Ott üldögéltem hát a váróban, Summers irodája előtt. Az Egyesült Államok volt pénzügyminisztere, a Harvard volt rektora, a legnagyobb közgazdasági díjak büszke birtokosa némi késéssel hívott be.

Először is elolvasta a titkárnője által kinyomtatott dolgozatomat arról, hogy milyen hatást gyakorolt a rasszizmus Obama választási eredményeire. Summers gyorsolvató. Miközben faldja az oldalakat, olykor kidugja és jobbra csúsztatja a nyelvét, a szeme pedig villámgyorsan jár balról jobbra, miközben egyre halad lefelé. Úgy olvas egy társadalomtudományi tanulmányt, ahogy a nagy zongoristák előadnak egy szonátát: olyan erősen koncentrálnak, hogy a külvilág megszűnik számára létezni. Alig öt perc alatt végzett a harmincoldalas dolgozatommal, majd így szólt:

„Szóval azt állítja, hogy a »nigger« szóra irányuló Google-keresések a rasszizmusra utalnak. Ebben van ráció. Megmagyarázza, hogy hol volt kisebb a támogatottsága Obamának, mint Kerrynek. Ez érdekes. De amúgy tényleg egyformának tekinthetjük Obamát és Kerryt?”

En meg így feleltem: „A politikatudósok úgy értékelik, hogy ideológiailag közel állnak egymáshoz. Ráadásul nincs korreláció a rasszismus és a kongresszusi választások szavazati arányának változása között. Még akkor sem, ha figyelembe vesszük a szociológiai összetételt, a templomba járást és a fegyverbirtoklást.” Így társalognak egymással a közgazdászok. Egészen felvillanyozódtam.

Summers szótlanul méregetett, majd a tévé felé fordult, amely a CNBC-re volt állítva. Utána megint engem méregetett egy sort, majd megint a tévét nézte. Aztán ismét hozzám fordult. „Oké, tetszik a tanulmánya” – mondta. – „Min dolgozik még?”

A következő hatvan perc valószínűleg életem legpezsdítőbb intellektuális élménye volt. Beszélgettünk kamatokról és inflációról, rendvédelemről és bűnözésről, üzletről és jótékonykodásról. Azok, akik abban a szerencsében részesülnek, hogy alkalmuk nyílik találkozni Summersszel, nem véletlenül vannak annyira lenyűgözve tőle. Számos hihetetlenül okos emberrel volt már szerencsém szót váltani életem során; Summers volt közülük a legokosabb. Szinte megszállottja a gondolatoknak, és ez sokszor bajba is sodorja. A Harvard rektori székéből például azért kellett felállnia, mert felvetette, hogy a természettudományos pályákon talán azért van kevés nő, mert a férfiak nagyobb változatosságot mutatnak IQ tekintetében. Summers már csak ilyen: ha egy gondolatot érdekesnek talál, akkor ki mondja, még akkor is, ha egyeseknek sérti a fülét.

A találkozónk ekkorra már fél órával túlnyúlt a tervezett kereten. Vérpezsdítően izgalmas volt a beszélgetés, de még mindig gőzöm sem volt róla, hogy miért vagyok ott, ahogy arról sem, hogy mikor kellene távoznom, valamint arról sem, hogy ha távoznom kell, azt miből fogom meg-

tudni. De volt egy sanda gyanúm, hogy talán már maga Summers sem tudja, miért akart találkozni velem.

És akkor feltette az egymillió, vagy talán inkább egymilliárd dolláros kérdést: „Mit gondol, előre tudják jelezni az adatai a tőzsdepiac alakulását?”

Aha! Végre kiderült, hogy miért hívatott az irodájába.

Korántsem Summers volt az első, aki ezt megkérdezte tőlem. Apám általában mindig is nagyon támogató volt a rendhagyó kutatásaimmal kapcsolatban. Egyszer aztán mégis belekérdezett a közepébe: „Rasszizmus, gyerekbántalmazás, abortusz... Mondd csak, pénzt is lehet keresni ezzel a te nagy szakértelmeddel?” Szóba hozták a témát más családtagok és a barátaim is. Aztán a munkatársaim, aztán idegenek az interneten. Mindenki arra volt kíváncsi, hogy ki tudom-e választani a Google-keresések vagy más Big Data-adathalmaz segítségével a megfelelő részvényeket. Most viszont az Egyesült Államok volt pénzügyminisztere kérdezte ugyanezt! Kezdt komolyra fordulni a dolog.

Képesek-e tehát a Big Data-adatforrások arra, hogy sikeresen megjósolják az értékpapírok mozgását? A rövid válasz az, hogy nem.

Az előző fejezetekben már szó esett a Big Data négy különleges képességéről. A mostani fejezet viszont kizárólag a Big Data korlátairól szól: arról, hogy mit nem tudunk megcsinálni vele, és helyenként arról is, hogy mit nem kellene megcsinálni vele. Kezdhethük akár azzal is, hogy elmesélem, miért nem sikerült kijátszanunk a tőzsdepiacot Summersszel.

A harmadik fejezetben már említettem, hogy az újszerű adatok akkor hoznak a legjobb eséllyel komoly eredményt, ha az adott szakterület kutatottsága alacsony szintű. Az élet sajnálatos, de nagy igazsága, hogy sokkal könnyebb

újszerű felismerésekre jutni a rasszizmus, a gyermekbántalmazás vagy az abortusz tárgykörében, mint egy cég teljesítményével kapcsolatban. Ez azért van, mert sokan már most is komoly erőforrásokat vetnek latba azért, hogy akár mégoly csekély előnyhöz jussanak az üzleti teljesítmény mérése terén. A pénzügyi szférában öldöklő verseny zajlik. Ez máris ellenünk szólt.

Summers, aki pedig nem arról híres, hogy áradozva dicséri mások intelligenciáját, biztosra vette, hogy a hedge fundok jóval előttünk járnak. A beszélgetés során nagy hatással volt rám, milyen tisztelettel beszél róluk, hogy hány-szor utasította el azzal egy-egy felvetésemet, hogy szerinte arra már ők is rég rájöttek. Büszkén megmutattam neki például egy saját fejlesztésű algoritmust, amely lehetővé teszi, hogy teljesebb körű Google Trends-adatokat töltsék le. Azt mondta, ügyes. Amikor pedig rákérdeztem, hogy szerinte a Renaissance nevű kvantitatív hedge fund is rájött-e már erre az algoritmusra, elkuncogta magát, és így felelt: „Hát persze hogy rájöttek.”

De az, hogy lépést tudjunk tartani a hedge fundokkal, nem az egyetlen problémánk volt Summersszel, amikor új, nagy adathalmazokkal próbáltuk magunk alá gyűrni a piacokat.

A DIMENZIONALITÁS ÁTKA

Tegyük fel, hogy azt a stratégiát alkalmazzuk a tőzsdepiac mozgásának megjóslására, hogy keresünk egy szerencsepénzt – de olyat, amelyet gondos tesztelés útján választunk ki. Íme, a módszer: megszámozunk ezer darab pénzérmét egytől ezerig, majd két éven át minden reggel feldobjuk őket. Feljegyezzük, hogy melyik lett fej, melyik írás, va-

lamint azt is feljegyezzük, hogy aznap felfelé vagy lefelé mozdult-e el a Standard & Poor indexe. A végén átböngésszük az adatokat, és *voilà!* Már találtunk is valamit: az olyan napokon, amikor a 391-es érme fejjel felfelé esett le, az esetek 70,3 százalékában nőtt az S&P Index! Az összefüggés statisztikailag szignifikáns, nem is kicsit. Megvan hát a szerencsepénzünk!

Most már nincs más dolgunk, mint minden reggel feldobni a 391-es érmét, és ha fej, akkor részvényt vásárolni. A turis pólók és zacskós tésztalevesek kora lejárt. Jegyet váltottunk az aranyéletbe, éljen a 391-es!

Vagy nem.

Mi is áldozatul estünk „a dimenzionalitás átka” egyik legördögibb vetületének. Ez az átok bármikor lesújthat ránk, ha kellően sok változónk (vagyis „dimenziónk”) van – ebben az esetben ezer darab pénzérme –, megfigyelésünk viszont nem olyan rengeteg – ebben az esetben két év alatt 504 tőzsdenap. Ilyenkor a sok dimenzió egyike – ebben az esetben a 391-es érme – nagy valószínűséggel bizonyul szerencsepénznek. Ha azonban csökkentjük a változók számát, például csak száz érmét dobunk fel reggelente, sokkal kisebb eséllyel bizonyul bármelyik is szerencsepénznek. Ha pedig növeljük a megfigyelések számát, például húsz évig próbálkozunk az S&P Index mozgásának előrejelzésével, akkor az érmék nemigen fogják tudni tartani a lépést.

A dimenzionalitás átka a Big Data egyik legfőbb problémája, mivel az újfajta adathalmazokban sokszor hatványozottan több a változó, mint a hagyományos adatforrásokban. Változónak minősül minden egyes keresőkifejezés, minden egyes Twitter-kategória, stb. Aki azt állítja, hogy valamilyen Big Data-forrást felhasználva előre jelzi a pia-

ci mozgásokat, azzal sokszor csupán csak az történt, hogy csapdába ejtette az átok. Annyiról van szó, hogy megtalálta a 391-es érmének megfelelő tényezőt.

Jó példa erre az Indianai Egyetem és a Manchesteri Egyetem informatikusaiból álló kutatócsoport, amely azt állította, hogy az emberek Twitter-bejegyzései alapján előre lehet jelezni, merre mozdul el a piac. Kidolgoztak egy algoritmust, amely a tweetek alapján kódolni tudta a világ aznapi hangulatát. A harmadik fejezetben tárgyalt érzélelemlemzéshez hasonló technikákkal dolgoztak. Ők azonban nemcsak egyféle hangulatot kódoltak, hanem sokfélé: boldogságot, dühöt, kedvességet és így tovább. Arra jutottak, hogy a nyugalmat sugalló tweetek (például: „Most nyugis vagyok”) túlsúlya azt jelzi előre, hogy a Dow Jones Ipari Átlag hat nappal később. Létrehoztak hát egy hedge fund alapot, hogy pénzre váltsák az eredményeiket.

Mi volt a probléma?

Alapvetően az, hogy túl sok mindent teszteltek. Márpedig ha az ember elég sok dolgot tesztl, akkor ezek egyike a pusztán véletlenből kifolyólag statisztikailag szignifikáns lesz. Rengeteg érzelmet teszteltek. Ráadásul minden egyes érzelmet leteszteltek egy nappal, két nappal, három nappal, négy nappal, öt nappal, hat nappal és hét nappal a megjósolni kívánt tőzsdepiaci mozgás előtt. És ezt a rengeteg változót arra próbálták használni, hogy megmagyarázzák a Dow Jones néhány havi ingadozását.

A hat nappal korábbi nyugalom nem bizonyult érvényes indikátornak a tőzsdepiaci mozgások előrejelzésére. Esetükben a hat nappal korábbi nyugalom volt a mi hipotetikus 391-es érmeünk Big Data-alapú megfelelője. A Twitter-alapú hedge fundot pedig egy hónappal az alapítása után megszüntették, mivel a hozam igencsak haloványnak bizonyult.

De nem csak a tweetek alapján jóslógató hedge fundok küzdenek a dimenzionalitás átkával. Hanem az a számos természettudós is, akik arra keresik a választ a genetika segítségével, hogy kik is vagyunk.

Hála a Humán Genom Projektnek, ma már összegyűjthető és elemezhető az emberek teljes DNS-állománya. Úgy tűnt, hogy a projektben óriási lehetőségek rejlenek.

Lehet, hogy megtaláljuk a skizofréniát okozó gént? Lehet, hogy megtaláljuk az Alzheimer, a Parkinson és az amiotrófiás laterálszklerózist okozó gént? Lehet, hogy megtaláljuk – nagy nyelés – az intelligenciáért felelős gént? Lehet, hogy van egy gén, amely egy rakás pluszpontot hoz az IQ-teszten? Lehet, hogy van egy gén, amely zsenit csinál az emberből?

Robert Plomin, a jeles viselkedésgenetikus 1998-ban bejelentette, hogy megválaszolta a nagy kérdést. Plomin hozzájutott egy több száz diák DNS-ét és IQ-ját tartalmazó adathalmazhoz. Összevetette a „zsenik”, azaz a legalább 160-as IQ-val rendelkező diákok DNS-ét az átlagos IQ-jú társaikkal.

A két csoport DNS-e között döbbenetes mértékű eltérést talált. A különbség a 6-os kromoszóma egy eldugott zugában búj meg egy kevésbé ismert, de igen erős gén formájában, amely az agy anyagcseréjében játszik szerepet. Plomin megállapította, hogy a gén egyik verziója, az IGF2r a „zseniknél” kétszer olyan gyakran fordul elő.

„Bejelentették, hogy megvan az első gén, amely a magas fokú intelligenciához köthető” – harsogta a *The New York Times* szalagcíme.

Rögtön eszünkbe juthat, hogy Plomin kutatási eredménye számos etikai problémát felvet. Például azt, hogy a szülők végeztethetnek-e IGF2r-szűrővizsgálatot a születendő gyermekükön. Abortáltathatják-e a magzatot, ha várha-

tóan alacsony lesz az IQ-ja? Módosítsuk-e genetikailag az embereket, hogy magasabb legyen az IQ-juk? Korrelál-e az IGF2r a faji hovatartozással? Tudni akarjuk-e a választ erre a kérdésre? Folytatódjanak-e az IQ genetikájával kapcsolatos kutatások?

De mielőtt a bioetikusok nekiláttak volna, hogy megbirkózzanak ezekkel a fogas kérdésekkel, a genetikusok, köztük maga Plomin, egy sokkal alapvetőbb problémával szembesültek: azzal, hogy vajon pontos-e az eredmény. Valóban igaz-e, hogy az IGF2r előre tudja jelezni az IQ-t? Valóban igaz-e, hogy a zsenik nagy valószínűséggel hordozzák a gén valamelyik variánsát?

Hát nem. Történt, hogy Plomin néhány évvel az eredeti vizsgálat után hozzájutott egy másik mintához, amely szintén emberek DNS-ét és IQ-pontszámait tartalmazta. És az IGF2r ezúttal nem korrelált az IQ-val. Plomin visszavonta az állítását – amivel egyébként azt is bizonyította, hogy valódi tudós.

Ami azt illeti, az IQ genetikájával kapcsolatos kutatásokra általában is jellemző ez a mintázat: a tudósok bejelentik, hogy találtak egy génvariánst, amely előre jelzi az IQ-t, majd új adatok birtokába jutnak, és rájönnek, hogy az állításuk téves volt.

Példának okáért nemrégiben jelent meg egy tanulmány egy kutatásról, amelynek során egy Christopher Chabris által vezetett kutatócsoport tizenkét, az IQ-hoz köthető génvariánsokra vonatkozó, fontos bejelentést vizsgált meg. Tízezer ember adatait vizsgálták – és a tizenkét állítás közül egy sem igazolódott be: egyszer sem tudták reprodukálni a korrelációt.

Vajon mi lehet a gond ezekkel az állításokkal? A dimenzionalitás átka. Az emberek génállománya, amint azt a tu-

dósok ma már tudják, millióféle módon különbözik egymástól. Egészen egyszerűen túl sok a vizsgálandó gén.

Ha elég sok tweetet tesztelünk, akkor valamelyik a pusztán véletlen folytán korrelálni fog a tőzsdepiac mozgásával. Ha elég sok genetikai variánst tesztelünk, akkor valamelyik a pusztán véletlen folytán korrelálni fog az IQ-val.

Hogyan küzdhetnénk le a dimenzionalitás átkát? Fontos, hogy alázattal végezzük a munkánkat, és ne szeressünk bele az eredményeinkbe, hanem vessük alá őket további teszteknek. Például mielőtt feltennénk minden megtakarításunkat a 391-es érmére, érdemes megnézni, hogyan teljesít a következő néhány évben. A társadalomtudósok ezt úgy hívják, hogy „mintán kívüli teszt”. És minél több változóval próbálkozunk, annál alázatosabbnak kell lennünk. Minél több változóval próbálkozunk, annál szigorúbbnak kell lennie a mintán kívüli tesztnek. Nagyon fontos az is, hogy minden egyes tesztünket nyomon kövessük. Így pontosan megítélhetjük, mekkora a valószínűsége annak, hogy lesújt ránk az átok, és mennyire kell szkeptikusan viszonyulnunk az eredményeinkhez. Most pedig visszakanyarodhatunk Larry Summershez és hozzám. Elmesélem, hogyan próbáltuk kijátszani a tőzsdepiacot.

Summers első ötlete az volt, hogy jósoljuk meg a Google-keresések segítségével egyes kulcsfontosságú termékek, például az iPhone-ok jövőbeni eladási mutatóit, ami talán felfedheti előttünk egy-egy cég, például az Apple részvényeinek jövőbeni ármozgását. Az iPhone-ra való rákérések, illetve az iPhone-eladási mutatók között valóban volt korreláció. Ha tehát az emberek sokszor keresnek rá az iPhone-ra, akkor nagy összegben fogadhatunk rá, hogy mostanában szépen fogy az iPhone. Csakhogy ezt az információt már beépítették az Apple-részvények árába. Nyil-

vánvaló, hogy a hedge fundok is rájöttek arra, hogy ha az iPhone-ra sokszor keresnek rá, akkor az iPhone kelendő áru. Hogy a Google-keresések adataiból jöttek rá, vagy más forrásból, az már nem is számít.

Summers következő ötlete az volt, hogy jelezzük előre a fejlődő országokban eszközölt befektetéseket. Ha sok befektető önti a pénzt például a brazil vagy a mexikói gazdaságba, akkor az ottani cégek részvényei nyilván drágábbak lesznek. Tegyük hát egy próbát, hátha meg tudjuk jósolni a befektetési volumen növekedését az olyan Google-keresések alapján, mint „befektetés Mexikóban” vagy „befektetési lehetőségek Brazíliában”. Ez azonban zsákutcának bizonyult. Mi volt a gond? Az, hogy az ilyen keresések ritkák. Így az adatokból nem rajzolódott ki értelmezhető mintázat, csak egy nagy összevisszaság.

Ezt követően az egyes cégek részvényeire irányuló keresésekkel próbálkoztunk. Ha valaki rákeres arra, hogy „GOOG”, az talán jelentheti azt, hogy az illető Google-részvényt akar vásárolni. És valóban úgy tűnt, hogy az ilyen keresések előre jelzik a részvény intenzív forgalmát. Azt azonban már nem jelzik előre, hogy emelkedni fog-e az ára, vagy csökkenni. Komoly akadály, hogy a keresések-ből nem derül ki, hogy venni akar az illető, vagy eladni.

Egy szép napon izgatottan megmutattam Summersnek az új ötletemet: úgy tűnt, hogy az „aranyat venni” keresőszóval indított keresések korreláltak az arany árának emelkedésével. Summers azt mondta, teszteljem a dolgot a következő időszakra is, hogy lássuk, továbbra is fennáll-e a korreláció. És nem állt fenn, talán azért, mert valamelyik hedge fund is rájött erre az összefüggésre.

Végezetül, jó néhány hónap elteltével a tesztjeinkből nem derült ki semmi használható. Ha végignéztük volna

a Google-keresőkifejezések milliárdjait, nyilván találtunk volna olyat, amely, ha csak gyengén is, de korrelál a tőzsdépiaci mutatókkal. De ez alighanem csak a mi 391-es érménk lett volna.

A MÉRHEŐSÉG TŰLHANGSŰLYOZÁSA

A connecticuti New Havenben Zoë Chance, a Yale marketingprofesszora 2012 márciusában kivett az irodai postaládájából egy kis fehér lépésszámlálót. Azért rendelte a készüléket, mert meg akarta nézni, hogy ösztönözhet-e valakit több testmozgásra egy szerkezet, amely megszámlolja, hányat lép egy nap, majd az eredmény alapján pontozza.

Ami ezután történt, az szabályos „Big Data-rémálom” volt, amint azt Chance TEDx-előadásából is megtudhatjuk. Mint elmondta, annyira rögeszméjévé vált a lépésszám növelése, hogy egyfolytában gyalogolt: a konyhából a nappaliba, onnan az ebédlőbe, majd a pincébe, aztán a dolgozószobába. Gyalogolt kora reggel, késő este, gyalogolt a nap szinte minden órájában: húszezer lépés minden huszonegy órában. Naponta több százszor ránézett a készülékre, és az emberi kommunikációja nagyrészt arra redukálódott, hogy más lépésszámláló-tulajdonosokkal csevegett online, akikkel megvitatták a lépésszámnövelő stratégiáikat. Mint felidézi, annyira a megszállottjává vált a lépésszámlálásnak, hogy még a hároméves kislányára is felszerelte a készüléket, amikor az mászkált.

Annyira a rögeszméjévé vált a kijelzőn megjelenő szám növelése, hogy többé egyáltalán nem volt képes józanul gondolkodni. Már el is felejtette, hogy a szám növelésének célja a testmozgás, nem pedig az, hogy a kislánya tegyen

pár lépést. A lépésszámlálóról szóló tudományos kutatásból sem lett semmi. Chance végül megszabadult a készüléktől, miután egy este elesett, amikor a végsőkéig kimerülve lépni akart még néhányat. Bár Chance hivatása szerint adatorientált kutató, az élmény igen nagy hatással volt rá. „Az eset kételyt ébresztett bennem azzal kapcsolatban, hogy biztosan mindig jó dolog-e, ha újabb adatokhoz nyílik hozzáférésünk” – mondta Zoë Chance.

Elég szélsőséges történet. Mégis rámutat egy potenciális problémára azzal kapcsolatban, ha az ember adatok alapján akar döntéseket hozni. A számok bódítóak tudnak lenni. A megszállottjaivá tudunk válni, és közben szem elől téveszthetjük a lényegesebb megfontolásokat. Zoë Chance például többé-kevésbé az egész lépésszámláláson kívüli életét szem elől tévesztette.

De már annak is lehet hátulütője, ha nem ennyire szélsőséges módon, rögeszmésen habarodunk bele a számokba. Gondoljunk például arra, mekkora súlyt fektetnek a huszonegyedik századi amerikai iskolákban a tesztekre – és mennyire az alapján ítélik meg a tanárokat, hogy hány pontot érnek el a diákjaik. Miközben jogos igény, hogy legyen objektív mércéje annak, ami a tanteremben történik, sok olyan is történik ott, amit nem könnyű számszerűsíteni. Ráadásul ez a tesztmánia sok tanárt arra készítetett, hogy magát a tanítást is pusztán a tesztekre való felkészítésként fogja fel – sőt olyan is akad, aki ennél is tovább megy. Brian Jacob és Steven Levitt tanulmánya kimutatta, hogy egyes pedagógusok egyenesen csálnak a tesztekkel.

A probléma a következő: azok a dolgok, amiket meg tudunk mérni, nem mindig pont ugyanazok, amik igazán fontosak lennének a számunkra. Könnyű megmérni, hogyan boldogulnak a diákok a feleletválasztós tesztekkel.

Egyáltalán nem könnyű azonban megmérni a kritikai gondolkodást, a kíváncsiságot vagy az egyéni fejlődést. Azaz, hogy megpróbálunk növelni egyetlen, könnyen mérhető számot – például a teszten elért pontszámot vagy a napi lépésszámot –, nem mindig jutunk közelebb ahhoz, amit valójában el akarunk érni.

A weboldalát fejleszteni igyekvő Facebook is szembesül ezzel a veszéllyel. A cég tonnaszám rendelkezik adatokkal arról, hogy miként használják az emberek a Facebookot. Mi sem egyszerűbb, mint megnézni, hányan lájkoltak egy bizonyos hírfolyamsztorit, hányan kattintottak rá, hányan kommentelték, hányan osztották meg. Alex Peysakhovich, a Facebook adattudósa szerint azonban, akivel leveleztünk a témában, egyik mutató sem helyettesíti tökéletesen a fontosabb kérdéseket: milyen volt az oldal által nyújtott felhasználói élmény? Összehozta-e a sztori a felhasználót a barátaival? Tájékoztatta-e a világ dolgairól? Tudott-e rajta nevetni?

Jó példa a kilencvenes években lezajlott baseball-adatforradalom is. Akkoriban sok csapat rákapott az egyre kifinomultabb statisztikákra: inkább ezek alapján hoztak döntéseket, mintsem az ódivatú játékosmegfigyelés alapján. A bökkenő az volt, hogy csak a támadójátékot és a dobóteljesítményt könnyű mérni, a mezőnyjátékot viszont nem. Ezért egyes klubok alábecsülték a védekezés fontosságát. Nate Silver *The Signal and the Noise (A jel és a zaj)* című könyvében úgy becsüli, hogy az Oakland Athletics adatorientált csapata (őket mutatja be a *Pénzcsináló* című baseballfilm) a kilencvenes évek közepén évi nyolc-tíz győzelmet engedett ki a kezéből a pocsék védekezése miatt.

A megoldás nem mindig a még több Big Data. Ahhoz, hogy a Big Data a lehető legjobban működjön, sokszor fű-

szerezni kell egy kis emberi ítélőképességgel, valamint kis elemszámú felmérésekkel, amelyeket ehelyütt hívjunk talán „small datának”. Amikor Nate Silver interjút készített Billy Beane-nel, az Oakland Athletics klubigazgatójával, a *Pénzcsináló* főhősével, Beane elmondta, hogy máris megnövelte a játékosmegfigyelésre szánt költségkeretet.

Ahhoz, hogy kitöltse az óriás adathalmaz hézagjait, a Facebooknak is a régi, ódivatú módszerhez kell folyamodnia: meg kell kérdeznie az emberektől, hogy mit gondolnak. Ezért a hírfolyamot letöltő facebookozók százainak tesznek fel kérdéseket az ott megjelenő sztorikról. A lájkok, kattintások, kommentek automatikusan gyűjtött adathalmazai tehát kis adatokkal egészülnek ki („Akarod látni ezt a posztot a hírfolyamodban?” „Miért?”). Bizony, olykor még egy ilyen látványosan sikeres „Big Data-s” cégnek is kapóra jön egy olyan információforrás, amelyet amúgy eddig többnyire eléggé leszóltam ebben a könyvben: a kis elemszámú felmérés.

Ami azt illeti, a „small data” mint a fő adatbázisokat – kattintások, lájkok és posztok óriási gyűjteményét – kiegészítő adatforrásra való igény okán a Facebook adattudóscsapata egészen máshogy néz ki, mint amilyennek valószínűleg elképzelik. A Facebook szociálpszichológusokat, antropológusokat, szociológusokat is alkalmaz, pontosan azért, hogy felfedjék azt, amit a számok elhallgatnak előlünk.

Egyes oktatásügyi szakemberek is egyre inkább felfigyelnek a Big Data vakfoltjaira. Egyre szélesebb körű, országos szintű erőfeszítések irányulnak arra, hogy a tömeges tesztelés kis adatokkal egészüljön ki. Elszaporodtak a diákok, illetve a szülők körében végzett felmérések, valamint a tanárok megfigyelése is, vagyis amikor tapasztalt pedagógusok látogatják a kollégáik óráit.

„A tankerületek kezdenek rájönni, hogy nem lenne szabad kizárólag a teszteredményekre koncentrálniuk” – véli Thomas Kane, a Harvard neveléstudomány-professzora. A Bill & Melinda Gates Alapítvány hároméves kutatása is alátámasztja, hogy az oktatásügyben a nagy adat és a kis adat is egyaránt értékes. A tanulmány szerzői azt elemezték, hogy a tesztpontszámon alapuló modellek, a diákok körében végzett felmérések vagy a tanármegfigyelések mérik-e a legjobban, hogy mely pedagógusok fejlesztették a legjobban a diákok tudását. A legjobb eredményeket akkor kapták, amikor a háromféle módszert egy kombinált pontrendszerben vegyítették. „Mindegyik mérési módszer hozzátesz valami értékeset a kutatáshoz” – állapítják meg a jelentés zárszavában.

Ami azt illeti, én is épp akkoriban találkoztam Jeff Sederrel a floridai Ocalában, amikor kiderült számomra, hogy sok Big Data-s szervezet kis adatokkal tömi be a lyukakat. Tudják, Seder az a Harvardon végzett lóguru, aki egy óriási adathalmaz tanulságai alapján megjósolta American Pharoah sikerét.

Seder, miután megosztotta velem a számítógépen tárolt adatait, és matekoztunk egy kicsit, beismerte, hogy van még egy fegyvere: Patty Murray.

Ez a Murray nevű nő Sederhez hasonlóan rendkívül intelligens, és szintén elit helyen szerzett diplomát: a pennsylvaniai Bryn Mawr lányfőiskolán. A vidéki élet kedvéért ő is otthagyta New Yorkot. „A lovakat jobban szeretem, mint az embereket” – vallja be. Az ő lófélértékelő módszere azonban némileg tradicionálisabb, mint a Sederé. Úgy jár el, mint a lőügynökök többsége: személyesen vizsgálja meg a paripákat. Szemrevételezi a járásukat, megnézi, van-e rajtuk sebhely, horzsolás, és kifaggatja a tulajdonosokat.

Ezután Sederrel közösen választják ki a megvásárlásra ajánlott lovakat. Patty Murray kiszimatolja az olyan problémákat, amelyekre Seder adatai alapján nem derül fény, hiába az övé minden idők leginnovatívabb és legfontosabb lóadathalmaza.

Azt jóslom, hogy a Big Data-revelációk revolúcióhoz fognak vezetni. De ez nem azt jelenti, hogy mostantól elég, ha a felmerülő kérdéseket jól megdobáljuk adatokkal. A Big Data nem teszi feleslegessé a világ megértését szolgáló egyéb módszereket, amelyeket az emberiség az évezredek során kifejlesztett. Ezek kiegészítik egymást.

8. fejezet

Nagy adat, nagy gond? Amit nem kellene

A Big Data óriási ereje olykor annyira lenyűgöző, hogy egyenesen ijesztő. Etikai kérdéseket vet fel.

A TÚLHATALMÚ CÉGEK VESZÉLYE

A közelmúltban három közgazdász (Oded Netzer és Alain Lemaire a Columbiáról, valamint Michal Herzenstein a Delaware-i Egyetemről) azt vizsgálta, miként lehetne előre jelezni, hogy mekkora eséllyel fizeti vissza a hitelfelvevő a banktól kapott összeget. A tudósok egy személyközi hitelezéssel foglalkozó weboldal, a Prosper adatait használták fel. A Prosperen a potenciális hitelfelvevők röviden leírják, miért van szükségük a kölcsönre, és miért tudják nagy valószínűséggel visszafizetni, a potenciális hitelezők pedig eldöntik, hogy adnak-e nekik pénzt. Összességében a hitelek mintegy 13 százaléka dől be.

Kiderült, hogy a potenciális hitelfelvevők szóhasználata erős indikátora a törlesztés valószínűségének. Még akkor is fontos ez az indikátor, ha figyelembe veszünk más releváns információkat is, amelyeket a hitelezők meg tudtak szerezni a potenciális hitelfelvevőkről, például a hitelértékelésüket, illetve a jövedelmüket.

Az alábbi listán tíz olyan kifejezés szerepel, amelyek a kutatás szerint gyakran előfordulnak a hiteligénylésekben. A tízből öt pozitívan korrelál a hitel visszafizetésével, a másik öt pedig negatívan. Azaz ötöt rendszerint olyan emberek használnak, akikben megbízhatunk, a másik ötöt pedig olyanok, akikben nem. Lássuk, ki tudja-e találni, melyik kifejezés melyik csoportba tartozik!

Isten	alacsonyabb kamat	adózás utáni
ígérem	fizetni fogok	kórház
tehermentes	diplomás	
minimális törlesztőrészlet	köszönöm	

Azt gondolhatnánk – vagy legalábbis remélhetnénk –, hogy ha egy udvarias, nyíltan vallásos ember a szavát adja rá, akkor felettebb valószínű, hogy vissza is fogja fizetni a hitelt. De valójában nem ez a helyzet. Az adatok azt mutatják, hogy az ilyen ember az átlagosnál kisebb valószínűséggel fogja törleszteni az adósságát.

Íme, a tíz kifejezés a visszafizetés valószínűsége szerint csoportosítva:

AZOK ÁLTAL HASZNÁLT KIFEJEZÉSEK,
AKIK A LEGNAGYOBB ESÉLYEL FIZETIK VISSZA A HITELT

tehermentes	adózás utáni	diplomás
alacsonyabb kamat	minimális törlesztőrészlet	

AZOK ÁLTAL HASZNÁLT KIFEJEZÉSEK,
AKIK A LEGKISEBB ESÉLYEL FIZETIK VISSZA A HITELT

Isten	fizetni fogok	kórház
ígérem	köszönöm	

Mielőtt megvitátnánk a tanulmány etikai vonatkozásait, gondoljuk végig a szerzők segítségével, hogy mit árul el mindez az emberekről! Mit kezdünk a két különböző kategóriába sorolt szavakkal?

Először nézzük a visszafizetést valószínűsítő kifejezéseket. Az „alacsonyabb kamat” és az „adózás utáni” bizonyos fokú pénzügyi műveltséget jelez a hitelfelvevő részéről, így talán nem meglepő, hogy korrelálnak a visszafizetés nagyobb valószínűségével. Ha pedig valaki az elért eredményeiről beszél, azaz közli, hogy ő „diplomás” vagy „tehermentes”, az szintén valószínűsíti, hogy fizetni fog.

Most pedig lássuk a hiteltörlesztést nem valószínűsítő kifejezéseket! Általában véve elmondható, hogy ha valaki azt mondja, fizetni fog, akkor nem fog fizetni. Minél határozottabban ígéri, annál valószínűbb, hogy megszegi az ígéretét. Aki azt írja: „Ígérem, hogy fizetni fogok, Isten engem úgy segéljen!”, annál a legkisebb a pontos törlesztés esélye. Ha pedig valaki a könyörületünkre apellálva elmondja, hogy kell a pénz, mert valamelyik családtagja kórházban van, akkor szintén eléggé valószínűtlen, hogy viszontlátjuk a pénzünket. Sőt mi több, bármely családtag (férj, feleség, gyerek, apa, anya) említése annak a jele, hogy az illető nem fog fizetni. A nem teljesítés előjele továbbá a „magyaráz” ige használata, hiszen aki megpróbálja elmagyarázni, hogy miért lesz képes visszavizetni a kölcsönt, az valószínűleg nem lesz képes rá.

Hogy a „köszönöm” szó használata miért valószínűsíti a hitel bedőlését, arra nincs elméletük a szerzőknek.

Összegezve: a kutatók szerint, ha valaki részletezett tervet ad elő a törlesztésre vonatkozóan, és megemlíti a múltbeli, teljesített kötelezettségeit, az arra utal, hogy vissza fogja fizetni a hitelt. Ha valaki ígéretet, és a könyörü-

letességünkre apellál, az annak az egyértelmű jele, hogy a hitel be fog dőlni. Függetlenül az okoktól (vagy attól, hogy mit árul el az emberi természetről, hogy ha valaki megígér valamit, az annak a biztos jele, hogy valójában nem fogja megtenni), a tudósok arra jutottak, hogy a teszt rendkívül értékes információval szolgál, és alkalmas a problémás hitelek kiszűrésére. Aki kiejti a száján Isten nevét, az 2,2-szer nagyobb valószínűséggel bizonyul fizetéseképtelennek. Az „Isten” szó használata a nem teljesítés legerősebb indikátorai közé tartozott.

A szerzőknek azonban meggyőződésük, hogy a tanulmányuk etikai kérdéseket is felvet. Míg ők csupán egy tudományos célú tanulmányt készítettek, akadnak olyan cégek, amelyek a hitelbírálathoz bevallottan felhasználnak online adatokat. Vajon elfogadható ez? Akarunk olyan világban élni, ahol a cégek az általunk leírt szavakból következtetnek arra, hogy vissza fogjuk-e fizetni a hitelt? Ez legalábbis elég bizarr, de talán az sem túlzó megfogalmazás, hogy riasztó perspektíva.

Az is elképzelhető, hogy a hiteligénylőnek a közeljövőben már nemcsak a pénzügyi előélete, hanem az online tevékenysége miatt is lesz félnivalója. És az sem kizárt, hogy abszurdnak tűnő tényezők alapján ítélik meg majd felette – például annak alapján, hogy használja-e a „köszönöm” szót, és hivatkozik-e Isten nevére. Mi lesz továbbá azzal a nővel, aki valóban segíti a kórházban fekvő nővérét, és biztosra vehető, hogy rendszeren törleszteni fog? Borzasztóan hangzik, hogy neki azért kell majd bűnhődnie, mert azokról, akik azt állítják, hogy orvosi költségekre kell nekik a pénz, összességében gyakran derült ki, hogy hazudtak. Ha ez így működne, a világ félelmetesen disztópikus jelleget öltene.

Az etikai kérdés így hangzik: joguk van-e a vállalatoknak ahhoz, hogy a szolgáltatásaikhoz közvetlenül nem

kapcsolódó, absztrakt, de statisztikai értelemben jó előrejelző erővel bíró kritériumok alapján bírálják el a szolgáltatásaikra való alkalmasságunkat?

Hagyjuk most egy kicsit a pénzügyi világot, és nézzük meg, milyen következményekkel jár az efféle gyakorlat, mondjuk, a munkaerő-felvétel gyakorlatára nézve. A munkaadók ma már egyre többet kutakodnak a közösségi médiában a jelöltek kiválasztásakor. Ha ezt azzal a céllal tesszik, hogy kiszűrjék az olyanokat, akik rossz hírért keltik a korábbi munkaadóiknak, vagy szétkürtölik azok titkait, az talán nem etikátlan. Talán még az is igazolható bizonyos fokig, ha nem akarnak felvenni olyan embert, akinek Facebook- vagy Instagram-posztjai mértéktelen alkohelizálásról árulkodnak. De mi a helyzet akkor, ha egy ártalmatlannak tűnő indikátor korrelál valamivel, ami a munkaadó számára fontos?

A Cambridge-i Egyetem és a Microsoft kutatói 58 ezer amerikai facebookozót kértek meg különböző, személyiségükre, illetve intelligenciájukra vonatkozó tesztek kitöltésére. Az eredményekből azt szűrték le, hogy a Facebook-lájkok gyakran korrelálnak az IQ-val, az extroverzióval, valamint a lelkiismeretességgel. Aki például lájkolja a Facebookon Mozartot, a vihart és a csavart sült krumplit, annak jellemzően magasabb az IQ-ja. Aki a Harley-Davidson motorokat, a Lady Antebellum nevű countryegyüttest vagy az „I Love Being a Mom” (Imádok anyuka lenni) című weboldalt lájkolja, annak jellemzően alacsonyabb az IQ-ja. Lehet, hogy a korrelációk egy része a dimenzionalitás átkának tudható be. Ha elég sok mindent tesztlünk, néhány dolog véletlenszerűen korrelálni fog egymással. De az is lehet, hogy bizonyos érdeklődési körök valóban korrelálnak az IQ-val.

Ezzel együtt nem lenne fair, ha egy okos ember, aki természetesen szeret Harley-val furikázni, nem kaphatna a ké-

pességeinek megfelelő munkát, mert tudtán kívül azt jelezte, hogy alacsony az intelligenciaszintje.

Az igazat megvallva a probléma nem teljesen új. Az emberek felett már régóta ítéleznek olyan tényezők alapján, amelyek nem függenek össze közvetlenül a munkatelményükkel – a kézfogásuk erőssége vagy éppen az öltö-zékük rendezettsége alapján. Az adatforradalom egyik veszélye abban rejlik, hogy az életünk egyre inkább számszerűsíthető, így az efféle közvetett bizonyítékokra épülő értéktételek egyre rejtélyesebbé, miközben a kutakodás egyre tola-kodóbbá válik. A jobb hatásfokú előrejelzés kifinomultabb, egyúttal aljasabb diszkriminációhoz vezet.

A jobb minőségű adatok egy másfajta diszkriminációt is eredményezhetnek, amelyet a közgazdászok úgy hívnak, hogy ársziszakrimináció. A cégek sokat törnek a fejüket a termékeik, szolgáltatásaik árszabásán. Ideális esetben a maximumot szeretnék elkérni, vagyis annyit, amennyit a fogyasztók még hajlandók kifizetni. Így fogják elérni a lehető legnagyobb profitot.

A legtöbb cég végül egyféle ár mellett dönt, és azt fizet-teti meg mindenkivel. De vannak olyan esetek is, amikor a cég tudja, hogy egy bizonyos csoport tagjai átlagosan többet is megadnának az áruért. Ezért kerül többbe a keresőképességük csúcsán lévő, középkorú nézők mozijegye, mint a diákoké és a nyugdíjasoké. Ahogy a légitársaságok is ezért adják sokszor drágábban a last-minute jegyeket: „ársziszakriminálnak”.

A Big Data segítségével a cégek lényegesen hatékonyabban feltérképezhetik, mennyit is hajlandók fizetni a fogyasztók – és hatékonyabban is húzhatnak le egyes csoportokat. Az Optimal Decisions Group úttörő volt abban, ahogy az adattudomány eszközeit felhasználva igyekezett előre jelezni, mennyit hajlandók fizetni az emberek a biz-

tosításért. Hogy hogyan csinálták? Pontosan azt a módszert alkalmazták, amelyet már tárgyaltunk ebben a könyvben. Megkeresték azokat a korábbi ügyfeleiket, akik a legjobban hasonlítottak az újonnan biztosítást kötni szándékozókra – és megnézték, hogy az előbbiek mekkora biztosítási díjat voltak hajlandók magukra vállalni. Más szóval futtattak egy alteregókeresést. Az alteregókeresés igen szórakoztató tevékenység, ha azt akarjuk megjósolni, hogy újra a régi fényben fog-e tündökölni egy baseballjátékos. Az alteregókeresés egyenesen nagyszerű tevékenység, ha abban segít, hogy valaki meggyógyuljon. Ha viszont az alteregókeresés ahhoz járul hozzá, hogy egy nagyvállalat kisajtolja belőlünk az utolsó centet is, akkor azért már nem annyira menő dolog. Az én költséző hajlamú öcsém például joggal emelhetne panaszt, ha többet kérnének tőle valamiért, mint tőlem, a szűkmarkú bátyjától.

A szerencsejáték is olyan terület, ahol veszélyes lehet, ha az üzemeltető rá tud zoomolni a vendégekre. A nagy kaszinók az alteregókereséshez hasonló módszerrel próbálják kiismerni a játékosokat. A cél itt is a profitmaximalizálás – az, hogy lehetőleg minél több pénzünk vándoroljon át a széfjükbé.

A dolog a következőképpen működik. A kaszinók azt feltételezik, hogy minden egyes szerencsejátékosnak megvan a maga „fájdalomküszöbe”. Ez azt az összeget jelenti, amelynek elvesztése már eléggé megijeszti a játékost ahhoz, hogy hosszabb időre távol maradjon a kaszinótól. Tegyük fel például, hogy Helen fájdalomküszöbe 3000 dollár. Ez azt jelenti, hogy ha ennyit veszít, akkor a kaszinó is hetekre, vagy akár hónapokra elveszít egy kuncsaftot. Ha azonban Helen csak 2999 dollárt bukik, akkor nem lesz ugyan boldog – hiszen ki örül annak, ha pénzt veszít? –, de a dolog nem is fogja annyira kedvét szegni, hogy másnap este ne játsszon újra.

Képzelve magát egy pillanatra a kaszinó vezetőjének helyébe! Képzelve el, hogy megérkezett Helen, és már dobálja is a pénzt a félkarú rablóba! Mi az ön szempontjából az optimális végkifejlet? Nyilván azért fog drukkolni, hogy Helen a lehető legközelebb kerüljön a „fájdalomküszöbéhez”, de azért ne lépje át azt. Azt szeretné tehát, hogy 2999 dollárt veszítsen, éppen annyit, hogy ön szép profitot zsebeljen be, de nem is akkorát, hogy utána Helen hosszabb időre elmaradjon.

Hogy tudjuk ezt elérni? Nos, vannak rá módszerek, hogy amikor már elbukott egy bizonyos összeget, leállítsa Hent. Felajánlhat neki például egy ingyenétkezést. Ha elég csábító az ajánlat, otthagyja a nyerőgépet.

A fájdalomküszöb megbecslése azonban magában rejt egy igen komoly kihívást. Honnan tudhatnánk, hogy pontosan hol van Helen „fájdalomküszöbe”? Az itt a gond, hogy az embereknek egészen eltérő lehet a „fájdalomküszöbük”. Helené 3000 dollár. Johné talán 2000 dollár. Bené esetleg 26 ezer dollár. Ha tehát Helen 2000 dollár mínuszban van, amikor rábeszéli, hogy hagyja abba a játékot, akkor ön az asztalon hagyott 1000 dollár tiszta profitot. Ha azonban túl sokáig vár, és Helen eléri a 3000 dolláros „fájdalomküszöbét”, jó időre elveszíti őt. Ráadásul talán nem is igen akarja elárulni önnek a „fájdalomküszöbét”. Talán még ő maga sem tudja, hol a határ.

Akkor hát mi a teendő? Aki idáig eljutott a könyvben, az valószínűleg kitalálja, mi a megoldás. Segítségül kell hívni az adattudományt. Mindent meg kell tudnia a kuntsaftjaink egy csoportjáról, amit csak lehet: az életkorukat, a nemüket, az irányítószámukat, a „szerencsejáték-kórtörténetüket”. És a „kórtörténetükből”, azaz a korábbi nyeres-vesztésük, jövés-menésük alapján megbecsülhetjük a „fájdalomküszöbüket”.

Összeszedezeti hát Helenről mindezeket az információkat, és megkeresi a hozzá hasonló játékosokat, vagyis többé-kevésbé az alteregóit. Aztán megnézi, hogy mi az ő túréshatáruk. Valószínűleg addig bírja Helen is. A Harrah's nevű kaszinó legalábbis így dolgozik, méghozzá a Terabyte nevű Big Data-raktározó cég segítségével.

Scott Gnau, a Terabyte vezérigazgatója *Super Crunchers* (Szuper számcsócsálók) című kiváló könyvében elmagyarázza, mit csinál egy kaszinóvezető akkor, ha azt látja, hogy egy törzsvendég már közel jár a fájdalomküszöbéhez: „Odamennek hozzá, és azt mondják: »Látom, nehéz napja van. Tudom, hogy szereti a steakházunkat. Szeretném, ha most meghívná a kedves feleségét vacsorázni a ház költségére.«”

Ez első hallásra a nagylelkűség csimborasszója: ingyenvacsora a steakházban! De a kaszinóvezetőt valójában a saját haszna vezérli. Igyekszik leállítani a törzsvendéget, nehogy az túl sokat veszítsen, és utána hosszabb szünetet tartson. Vagyis a vezetőség szofisztikált adatelemzés módszerével próbálja elérni, hogy hosszú távon a lehető legtöbb pénzt sajtolhassa ki a vendégekből.

Joggal tarthatunk attól, hogy ha a kaszinók, a biztosító társaságok, a hitelintézetek és más vállalatok egyre hatékonyabban fogják tudni felhasználni az online adatokat, túl nagy hatalomra tesznek szert felettünk.

Másrészt viszont a Big Data a fogyasztóknak is tud segíteni abban, hogy keményen odacsaphassanak a túlszámlázó vagy ócska árut szállító cégeknek.

Fontos fegyvert adnak a kezükbe például az olyan weboldalak, mint a Yelp, amely éttermekről és más szolgáltatókról tesz közzé értékeléseket. Michael Luca, a Harvard közgazdásza tanulmányában nemrég kimutatta, hogy a cégek mennyire ki vannak szolgáltatva a Yelp-értékelések kényének-kedvének. Luca Washington állam bevételi

adataival vetette össze az értékeléseket, és megállapította, hogy ha egy étterem eggyel kevesebb csillagot kap a Yel-pen, 5–9 százalékkal csökken a bevétele.

A vállalatokkal szembeni küzdelemben az olyan árösszehasonlító oldalak is a fogyasztók segítségére vannak, mint a Kayak és a Booking.com. A *Lökonómia* című könyvben olvashatunk arról, hogy ha egy internetes oldal elkezd ki-kö-zölni, mennyi a határozott idejű életbiztosítás díja a külön-böző társaságoknál, a díjak összege drasztikusan zuhanni kezd. A fogyasztók ettől kezdve tudni fogják, melyik tár-saság árazza túl a biztosítást, és átmehetnek egy másikhoz. Hogy ennek köszönhetően mennyi pénzt takarítottak meg összesen a fogyasztók? Évi egymilliárd dollárt.

Az internetes adatokból tehát a cégek megtudhatják, hogy mely ügyfeleket érdemes elkerülniük, és mely ügyfeleket tudják jól lehúzni. A fogyasztók pedig azt tudhatják meg, hogy mely cégeket érdemes elkerülniük, és melyek akarják lehúzni őket. A Big Data eddig mindkét félnek segítségére volt a fogyasztók és a vállalatok harcában. Ügyelnünk kell rá, hogy a küzdelem tisztességes eszközökkel folytatódjon.

A TÚLHATALMÚ KORMÁNYZAT VESZÉLYE

Amikor a volt barátja felbukkant a születésnap bulin, Ad-riana Donato tudta, hogy a fiú teljesen bekattant. Tudta, hogy dühös. Tudta, hogy depresszióval küzd. Csak egyva-lamit nem tudott a húszéves, zoológia szakos egyetemista lány, amikor az exe elhívta kocsikázni: azt, hogy a fiú, a huszonkét éves James Stoneham már három hete a gyil-kosság lehetséges módzatait és az idevonatkozó jogszabá-lyokat kutatja a neten, miközben olykor az exére, Adrianá-ra is rákeres.

Ha ezt is tudta volna, feltehetően nem száll be a kocsiba. És feltehetően nem késelik halálra aznap este.

A *Különvélemény* című filmben médiumok működnek együtt a rendőrséggel, hogy elejét vegyék a bűncselekmé-nyeknek. Lehet, hogy a valóságban a bűnüldöző szervek Big Data-hozzáférése lenne a megoldás? Figyelmeztetniük kellett volna Donatót, hogy a volt barátja vészjósló dol-gokat művel a neten? Ki kellett volna hallgatniuk Stone-hamet?

Először is el kell ismernünk: egyre több bizonyíték utal arra, hogy a bűntényekkel kapcsolatos Google-keresések korrelálnak az elkövetett bűntényekkel. Christine Ma-Kel-lams, Flora Or, Ji Hyun Baek és Ichiro Kawachi azt mutatta ki, hogy az öngyilkossággal kapcsolatos Google-keresések is erősen korrelálnak az állami szintű öngyilkossági rátá-val, mi pedig Evan Soltasszal azt mutattuk ki, hogy az isz-lamofób keresések (mint például a „gyűlölöm a muszlimo-kat” vagy a „muszlimokat ölni”) heti mennyisége korrelál az adott hét muszlimellenes gyűlöletbűncselekményeinek számával. Ha több ember ad hangot a kereséseiben annak, hogy valamit tenni akar, akkor azt a valamit többen is fog-ják megtenni.

Mit kezdjünk hát ezzel az információval? Íme, egy egy-szerű ötlet, amely talán nem vált ki túl sok vitát: a területi szintű adatoknak hasznát vehetjük az erőforrások elosztá-sában. Ha például egy városban elszaporodnak az öngyil-kossággal kapcsolatos keresések, akkor ott fokozhatjuk a megelőző tevékenységet. Mondjuk, társadalmi célú hir-detéseket készíttethet a város önkormányzata vagy a non-profit szervezetek, hogy az emberek lássák, hová fordul-hatnak segítségért. Ha pedig egy városban sokkal többen írják be, hogy „muszlimokat ölni”, mint korábban, akkor a rendőrség talán bölcsen teszi, ha felülvizsgálja a járőrözés

rendjét. Például több embert vezényelhetnek ki a helyi mecset védelmére.

Ám jobb, ha egy bizonyos lépéstől gondosan tartózkodunk: attól, hogy konkrét egyénekre szálljunk rá még azelőtt, hogy egyáltalán bármiféle bűncselekmény történt volna. Ez mindenekelőtt a privát szféra megsértése lenne. Etikai szempontból hatalmas távolság van aközött, hogy csak a kormány fér hozzá ezrek és százezrek keresési adataihoz, vagy a rendőrség is hozzáférhet egy adott személy keresési adataihoz. Etikai szempontból hatalmas távolság van aközött, hogy a helyi mecsetet védjük, vagy átkutatók valakinek a házáat. Etikai szempontból hatalmas távolság van aközött, hogy csak hirdetésekkel próbáljuk elejét venni az öngyilkosságoknak, vagy valakit akarata ellenére elmeorvosintézetbe zárunk.

És nemcsak etikai okokból kell roppant óvatosan bannunk az egyéni adatokkal, hanem adattudományi okokból is. Az adattudomány szempontjából is hatalmas távolság van aközött, hogy egy város tetteit kell-e megjósolni, vagy egyetlen személyét.

Egy pillanat erejéig térjünk vissza az öngyilkosságokhoz! Az Egyesült Államokban havonta mintegy 3,5 millió Google-keresés irányul erre a témára, és ezek többsége szuicid gondolatokat sejtet – úgymint „öngyilkos hajlam”, „öngyilkosság elkövetése” vagy „hogyan legyünk öngyilkosok”. Vagyis havonta több mint egy ilyen keresés jut minden száz amerikaiára. Erről egy Friedrich Nietzsche-idézet juthat eszünkbe: „Az öngyilkosságra való gondolás erős vigasz: számos gonosz éjszakán jut vele túl az ember.”* A Google keresési adatai igazolják, mennyire igaz ez, hogy mennyire

* Friedrich Nietzsche: *Túl jön és rosszon*. Ikon Kiadó, 1995, 63. Tatár György fordítása. (A szerk.)

általános az öngyilkosság gondolata. Mindazonáltal a tényleges öngyilkosságok száma a havi négyezret sem éri el az Egyesült Államokban. Az öngyilkosságról való képzelgés elképesztően gyakori. Az öngyilkosság nem az. Nem lenne hát sok értelme, ha a rendőrök mindenkihez bekopognának, aki valaha életében belekiabálta az online térbe, hogy szívesen kiloccsantaná a saját agyát – ha más miatt nem, hát már csak azért sem, mert a rendőrségnek semmi másra nem maradna ideje.

Vagy vegyük a vérszomjas iszlamofoób keresések példáját! 2015-ben úgy 12 ezerszer írták be a keresőbe Amerikában, hogy „muszlimokat ölni”, de összesen 12 olyan, muszlimok ellen elkövetett gyilkosság történt, amelyet utóbb gyűlöletbűncselekménynek minősítettek. Nyilvánvaló tehát, hogy a hátborzongató keresőkifejezést beírók túlnyomó többsége nem váltotta be a fenyegetését.

Egy kis matek segítségével megérthetjük, mi a különbség egy személy, illetve egy város viselkedésének előrejelzése között. Álljon itt egy egyszerű gondolat kísérlet. Tegyük fel, hogy egy egymillió lakosú városban egy mecset van. Tegyük fel továbbá, hogy ha valaki nem keres rá arra, hogy „muszlimokat ölni”, akkor csupán egy a százmillióhoz az esély arra, hogy megtámad egy mecsetet. Ha viszont valaki rákeres arra, hogy „muszlimokat ölni”, akkor a támadás esélye sokkal nagyobb lesz: egy a tízezerhez. Tegyük fel, hogy az iszlamofobia az egekbe szökik, és százról ezerre nő a „muszlimokat ölni” keresések száma.

Ez esetben a mecset megtámadásának matematikai esélye körülbelül az ötszörösére, két százalékról tíz százalékra nő. Annak esélye viszont, hogy egy olyan személy, aki rákeresett arra, hogy „muszlimokat ölni”, tényleg megtámadja a mecsetet, továbbra is egy a tízezerhez lesz.

Ilyen helyzetben nem az a helyes reakció, hogy bebör-

tönzünk mindenkit, aki rákeresett arra, hogy „muszli-mokat ölni”. Az se sokkal jobb, ha meglátogatják őket. Elenyésző az esélye annak, hogy közülük bárki is bűncselekményt kövessen el. Az viszont helyes reakció, ha védelem alá helyezzük a mecsetet, hiszen a megtámadásának a valószínűsége immár tíz százalék.

Nyilvánvaló tehát, hogy nagyon sok szörnyű keresés nem vezet szörnyű tettekhez.

Ezzel együtt, legalábbis elméletileg, lehetséges, hogy a keresések egyes típusai alapján elég nagy valószínűséggel megjósolhatók bizonyos szörnyű következmények. Lehetséges például, legalábbis elméletileg, hogy az adattudósok egyszer létrehoznak egy modellt, amely kiszúrhatta volna, hogy James Stoneham Adriana Donatóra vonatkozó keresései komoly aggodalomra adnak okot.

2014-ben mintegy 6000-szer kerestek rá pontosan arra a kifejezésre, hogy „hogyan öljük meg a barátnőnket”, és 400-an gyilkolták meg ténylegesen a barátnőjüket. Ha előzetesen az összes gyilkos rákeresett volna erre a kifejezésre, akkor elmondhatnánk, hogy az összes személy közül, aki rákeresett arra, hogy „hogyan öljük meg a barátnőnket”, minden tizenötödik be is váltotta a fenyegetését. Persze a barátnőjüket meggyilkolók jó része, sőt alighanem a túlnyomó többségük nem keresett rá pontosan erre a kifejezésre. Vagyis annak valószínűsége, hogy a „hogyan öljük meg a barátnőnket” kifejezésre való rákeresés gyilkossághoz vezet, nem 1:15, hanem annál kisebb, még hozzá alighanem sokkal kisebb.

De ha az adattudósok létrehoznának egy modellt, amely kimutatja, hogy egy konkrét személy meggyilkolásának esélye, mondjuk, 1:100, ezzel az információval talán már kezdenünk kellene valamit. A fenyegetett személynek legalább annyihoz biztosan joga lenne, hogy figyelmeztessék,

1:100 az esély arra, hogy egy bizonyos személy meg fogja ölni.

Összességében mégis érdemes óvatosnak lenni az ügyben, hogy keresési adatok alapján egyéni szinten próbáljuk megjósolni a bűncselekményeket. Az adatok egyértelműen arról árulkodnak, hogy a rengeteg szörnyű keresés csak ritkán vezet szörnyű tettekhez. És egyelőre nincs bizonyíték arra, hogy a kormány pusztán a keresések alapján, előre tudna jelezni egy szörnyű tettet. Nagyon óvatosnak kell tehát lennünk abban a kérdésben, hogy a kormány beavatkozhasson-e egyéni szinten a keresési adatokra alapozva. És nem is csupán etikai vagy jogi megfontolásból ajánlatos az óvatosság. Hanem, legalábbis egyelőre, adattudományi megfontolásból is, vagy valakit akarata ellenére elmeorvosintézetbe zárunk.:privát szférával kapcsolatos problémák, és a túlhatalmú kormányzat veszélye”.

Zárszó: hányan olvasnak végig egy könyvet?

Amikor aláírtam a szerződést erre a könyvre, tisztán láttam magam előtt, hogyan fog felépülni. Talán emlékeznek rá, hogy valahol az elején szó esett egy kis jelenetről, amely a családunk hálaadási vacsorája közben játszódott le. A szeretteim, ugyebár, kétségbe vonták az épelméjűségemet, és azon agyaltak, hogy vajon harminchárom éves koromra miért nem sikerült még megtalálnom az igazit.

A könyv zárszava tehát gyakorlatilag önmagát írta meg. Megtalálom az igazit, és feleségül veszem. Sőt mi több, a Big Datát fogom alkalmazni, hogy megismerkedjem vele! Esetleg itt-ott beleszőhetném a gondolatmenetbe az udvarlási folyamat apró-cseprő epizódjait. A történet pedig végül a zárszóban kerekedhetne ki, amely az esküvőm krónikája lehetne, egyúttal szerelmeslevél újdonsült nejemnek.

Ám az élet sajnos nem igazolta a várakozásaimat. Az, hogy könyvírás ürügyén bezárkóztam a lakásba, és elbújtam a világ elől, valószínűleg nem sokat segített a szerelmi életem fejlődésén. Így hát sajnos továbbra sem találtam meg életem párját. De ennél is fontosabb, hogy új zárszót kellett kitalálnom.

Átböngésztem a kedvenc könyveimet, hátha kiderül, mitől lesz igazán kiváló egy zárszó. Végül arra a konklúzióra jutottam, hogy a legjobb konklúziók felszínre hoznak valami fontosat, ami addig is jelen volt, de valahol a fel-

szín alatt lappangott. Ez a fontos dolog ennek a könyvnek az esetében a következő: a társadalomtudomány kezd igazi tudománnyá válni. És ez az újszerű, igazi tudomány arra készül, hogy jobba tegye az életünket.

A második rész elején szó esett Karl Popper Freud-kritikájáról. Popper, mint említettem, nem igazán tartotta tudományosnak Sigmund Freud hóbortos vízióját. Azt viszont nem említettem, hogy ez nem csupán Freud ellen intézett támadás volt, hanem sokkal általánosabb érvényű bírálat. Popper nem gondolta, hogy *bármely* akkori társadalomtudós különösebben tudományos dolgokat művelne. Egészen egyszerűen nem volt lenyűgözve attól, amit ezek az úgynevezett tudósok műveltek, mert úgy látta, hogy a módszereik nem éppen rigorózusak.

Mi motiválta Popper keresztes hadjáratát? Az, hogy korra legkiválóbb elmével – a legjobb fizikusokkal, a legjobb történészekkel, a legjobb pszichológusokkal – társalognva felfigyelt egy feltűnő különbségre. Amikor fizikusokkal beszélt, hitt abban, amit csinálnak. Persze néha ők is követtek el hibákat, és tudat alatti elfogultságaik olykor velük is a bolondját járatják. De a fizikusok akkor is egy olyan folyamat részesei voltak, amely egyértelműen mély igazságokat tárt fel a világról, és végül Einstein relativitáselméletében csúcsonyult ki. Ezzel szemben, amikor a világ leghíresebb társadalomtudósaival beszélgetett, Popper úgy érezte, hogy amit mondanak, az csak terjengős, fellengzős halandzsa.

Aligha Karl Popper az egyetlen, aki érzékelte ezt a különbséget. Abban nagyjából mindenki egyetért, hogy a fizikusok, a biológusok és a vegyészek valódi tudósok. Precíz kísérletek útján próbálják megfejtetni a fizikai világ működését. A közgazdászokat, szociológusokat, pszichológusokat viszont sokan „puha” tudósoknak tartják,

akik érthetetlen szakzsargonban zagyválnak össze mindenfélét, és ezért még fizetést is kapnak.

A Big Data-forradalom mindenképpen megváltoztatta ezt az általános nézetet, függetlenül attól, hogy mennyire volt egyáltalán megalapozott. Ha Karl Popper ma élne, és ellátogatna egy Raj Chetty, Jesse Shapiro, Esther Duflo, vagy akár (ennyi engedtessek meg) szerény személyem által tartott prezentációra, gyanítom, hogy nem az lenne a reakciója, mint annak idején. Sőt, őszintén szólva még az sem kizárt, hogy inkább, mondjuk, a húrelmélet lenne neki gyanús. Igen, felmerülhetne benne a kérdés, hogy vajon tudomány-e, amit az ezzel foglalkozó nagy koponyák csinálnak, vagy csak renyhe, öncélú agytorna.

Ha erőszakos filmet mutatnak be a városban, emelkedik a bűnözési ráta, vagy csökken? Ha egy hirdetés több emberhez jut el, akkor többen is vásárolják meg a hirdetett terméket? Ha húszéves korunkban bajnokságot nyer egy baseballcsapat, akkor valószínű, hogy negyvenévesen is nekik fogunk szurkolni? Ezek világos kérdések, amelyekre egyértelműen, igennel vagy nemmel felelhetünk. És a válaszokat is megtalálhatjuk a hiteles adatok tengerében.

Ez már tudomány, nem pedig áltudomány.

Ez persze nem jelenti azt, hogy a társadalomtudományi forradalom egyszerű, időtlen törvények formájában jön el.

Marvin Minsky, a néhai MIT-tudós, aki az elsők közt vizsgálta a mesterséges intelligencia lehetőségét, egyszer felvetette: elképzelhető, hogy a pszichológia vakvágányra tévedt, amikor a fizikát próbálta utánózni. A fizika abban sikeres, hogy egyszerű törvényszerűségeket tár fel, amelyek mindig, mindenhol érvényesek.

Nem biztos azonban, mondja Minsky, hogy az emberi agyra is ilyen törvények vonatkoznak. Valószínűbb, hogy az agyunk afféle „hackelés” bonyolult rendszere, amely-

nek részei egymás hibáit javítgatják. És hasonlóan komplex lehet a gazdasági, illetve politikai rendszer is.

Ennélfogva nem valószínű, hogy a társadalomtudományi forradalom takaros képletek formáját fogja öltetni, mint az $E=mc^2$ – és ha valaki mégis bejelentené, hogy itt az egyetlen takaros képletre alapozott társadalomtudományi forradalom, akkor ezt ajánlatos kételkedve fogadnunk.

A forradalom nem így fog beköszönteni, hanem lépésről lépésre, tanulmányról tanulmányra, eredményről eredményre. Vagyis úgy, hogy lassanként egyre jobban fogjuk érteni az emberi elme és a társadalom komplex rendszerét.

Egy rendes zárszó, azaz konklúzió nem csupán összegez, hanem előre, azaz újabb, később eljövendő dolgok felé is mutat.

Esetünkben ez igazán gyerekjáték. A könyvben tárgyalt adathalmazok forradalmi jelentőségűek, de javarészt feltáratlanok. Nagyon sok mindent tanulhatunk még belőlük, annál is inkább, mert, hogy őszinték legyünk, a tudósok többsége még alig-alig vett tudomást a digitális kor adat-robbanásáról. A világ legnevesebb szexológusai például ragaszkodnak a régi, jól bevált módszerhez. Néhány száz embert faggatnak a vágyaikról, a Pornhub és más hasonló oldalak adataira viszont nem kíváncsiak. A világ legnevesebb nyelvészei különálló szövegeket elemeznek, a könyvek millióiban kirajzolódó mintázatokra viszont nem kíváncsiak. A digitális forradalom jobbra érintetlenül hagyta a pszichológia, a politikatudomány és a szociológia egyetemen oktatott módszertanát is. Az adatrobbanás által megnyitott, nagyrészt feltáratlan területet meghagyták egy maroknyi haladó gondolkodású professzornak, rebellis doktorandusznak és hobbitudósnak.

Mindez meg fog változni.

Minden egyes ötletre, amelyről ebben a könyvben szót ejtettem, jut másik száz, amelyekkel ugyanilyen fontos lenne foglalkozni. Az általam tárgyalt kutatások csak a jég-hegy csúcsának csúcsát, a felszín felszínének kapirgálását jelentik.

Szóval mi jöhet még?

Például az, hogy radikálisan kibővül a minden idők leg-sikeresebb közegészségügyi tanulmányában alkalmazott módszertan. John Snow brit orvost a tizenkilencedik század közepén az foglalkoztatta, hogy mi okozza a londoni kolerajárványt.

Az a leleményes ötlete támadt, hogy a város térképén bejelölte az összes koleraesetet. Erre rájött, hogy a megbetegedések nagyrészt egy bizonyos szivattyúház körül csoportosultak, ami arra utalt, hogy a kolerát a baktériumtól fertőzött víz terjeszti. Így megdőlt az akkori idők közkeletű vélekedése, mely szerint a kórt a rossz levegő terjeszti.

A Big Data – illetve a Big Data által lehetővé tett „zoomolás” – megkönnyíti az efféle tanulmányok készítését. Ma már bármilyen betegséggel kapcsolatban kutathatunk a Google keresési adatai vagy más digitális egészségügyi adatok között. Megnézhetjük, vannak-e a világnak olyan apró zugai, ahol szokatlanul magas vagy szokatlanul alacsony az adott betegség előfordulási aránya. Utána pedig megnézhetjük, mi lehet a közös ezekben az apró zugokban. Van valami a levegőben? Vagy a vízben? Vagy a társadalmi normákban?

Megtehetjük ugyanezt a migrénnel. Megtehetjük ugyanezt a vesekővel. Megtehetjük ugyanezt a szorongással, a depresszióval, az Alzheimerrel, a hasnyálmirigyrákkal, a magas vérnyomással, a hátfájással, a székrekedéssel, az orrvérzéssel. Megtehetjük bármivel. Négyszázszor is elvégezhetjük azt

az elemzést, amelyet John Snow annak idején elvégzett (mire ezt írom, már el is kezdtem ezen dolgozni).

Ezt – vagyis hogy fogunk egy egyszerű módszert, és a Big Data segítségével rövid idő alatt több százszor elvégezzük ugyanazt az elemzést – bizvást nevezhetjük nagy léptékű tudománynak. Igen, a társadalomtudomány és a viselkedéstudomány határozottan kezdi kinőni magát, és egyre nagyobb léptékű lesz. A naggyá válást elősegíti, hogy rá tudunk zoomolni a különböző betegségekre. Egy további eszköz, amely segíthet a naggyá válásban: az A/B-tesztelés. Erről a módszerről már esett szó. Főként cégek élnek vele: így bírják rá az embereket, hogy kattintsanak a szalagcímekre és a hirdetésekre. De az A/B-tesztelés sokkal lényegesebb – és társadalmi szempontból értékeesebb – dolgok feltárására is alkalmazható, mint egy nyíl megalkotására, amely rábírja az embereket, hogy kattintsanak egy hirdetésre.

Benjamin F. Jones, a Northwestern Egyetem közgazdásza például arra használja az A/B-tesztelést, hogy így próbáljon segíteni a gyerekeknek a tanulásban. Jones közreműködött az EDU STAR nevű platform létrehozásában, amely lehetővé teszi, hogy az iskolák random módon teszteljék a különböző óravázlatokat.

Az oktatási szoftverek piacán számos cég képviselteti magát. Az EDU STAR úgy működik, hogy a diák bejelentkezik, és random módon különböző óravázlatokat mutatnak neki. Utána rövid teszteket tölt ki, hogy kiderüljön, mennyire sajátította el a tananyagot. Vagyis az iskolák megtudhatják, melyik szoftver válik be, azaz melyik érteti meg a legjobban az anyagot a diákkal.

Mint minden nagy A/B-tesztelő platform, az EDU STAR is produkál meglepő eredményeket. A pedagógusok nagyon izgalomba jöttek például egy óravázlattól, amely já-

tékokat használt fel arra, hogy könnyebben meg lehessen tanítani a diákoknak a törtszámokat. Ha játékos matek-órákat tartunk, akkor a diákok jobban élvezik a dolgot, könnyebben tanulnak, és a teszteken is jobban fognak teljesíteni. Igaz? Nem igaz. Azok a diákok, akik játékosan tanulták a törteket, rosszabb teszteredményeket produkáltak, mint a hagyományosabb módon okított társaik.

A gyerekeket rábírní, hogy többet tanuljanak, izgalmas és társadalmilag hasznos alkalmazása ennek a tesztmódszernek, amelyet a Szilícium-völgy eredetileg azért vetett be, hogy az emberek többször kattintsanak rá a hirdetésekre. Ahogy az is jó dolog, ha arra vesszük rá az embereket, hogy aludjanak többet.

Az átlagos amerikai 6,7 órát alszik éjszakánként. A többség ennél többet akar aludni, csak hát hamar eljön a 11 óra, és még mindig megy a tévében a napi sportösszefoglaló, vagy hív a YouTube. Így hát az ágnak várnia kell. A Jawbone nevű, hordozható kutyüket gyártó cég, amelynek több százezer ügyfele van, tesztek ezreit végzi, hogy rájöjjön, milyen beavatkozás segíthet a felhasználóknak abban, hogy azt tehessék, amire vágnak: hogy előbb lefeküdjének.

A Jawbone hatalmas diadalt aratott, méghozzá a kettős célkitűzés módszerével. Első lépésként arra kérték ügyfeleiket, hogy tűzzenek ki maguk elé egy nem túl nagyra törő célt. Küldtek nekik egy üzenetet, amely valahogy így szólt: „Úgy tűnik, hogy ön az utóbbi 3 nap alatt nem aludt túl sokat. Mi lenne, ha elhatározná, hogy ma 11:30-kor lefekszik? Tudjuk, hogy rendszerint 8-kor kel.” A felhasználók rákattinthatnak egy olyan opcióra, hogy „Benne vagyok”.

A második lépés az, hogy a Jawbone este 10:30-kor küld egy újabb üzenetet: „Közösen úgy döntöttünk, hogy ma megpróbál 11:30-kor lefeküdni. Most 10:30 van. Miért ne kezdené most?”

A Jawbone tesztje szerint ez a stratégia átlagosan 23 perc alvástöbbletet eredményezett. Azt nem sikerült elérniük, hogy ügyfeleik már 10:30-kor nyugovóra térjenek, de arra sikerült rávenniük őket, hogy az addigiakhoz képest korábban lefeküdjének.

Persze a stratégia minden elemét optimalizálni kell, és ehhez sok-sok kísérletre van szükség. Ha ugyanis az eredeti célkitűzést túl koraira löjük be, például azt kérjük az emberektől, hogy tűzzék ki célul a 11 órai lefekvést, akkor kevesen fognak belemenni a játékba. Ha viszont azt kérjük, hogy éjjélkor térjenek nyugovóra, azzal nem sokra megyünk.

A Jawbone A/B-tesztelés útján kereste a Google jobb felé mutató nyilának megfelelőjét. De itt ahelyett, hogy szereztek volna néhány új kattintást a Google hirdetési partnereinek, néhány perccel több pihenést szereztek a kimerült amerikaiaknak.

Tulajdonképpen nem kizárt, hogy a pszichológia mint tudományág kutatási eredményei drámaian javulhatnak, ha igénybe vennék a Szilícium-völgy eszköztárát. Epedkedve várom az első olyan pszichológiai tanulmányt, amely nem néhány egyetemistával végzett kísérletek tanulságait részletezi, hanem, mondjuk, ezer villámgyors A/B-teszt eredményét írja le.

Leáldozik annak a kornak, amikor a kutatók egyetlen teszthez hónapokon át toborozták a diákokat. Ehelyett a tudósok a digitális adatok segítségével alig pár másodperc alatt több száz, vagy több néhány ezer ötletet fognak letesztelni. Azaz sokkal kevesebb idő alatt sokkal többet fogunk megtudni.

A szöveg mint adat még sok mindenre meg fog tanítani minket. Hogyan terjednek a gondolatok? Hogy alakulnak ki új szavak? Hogy tűnnek el szavak? Hogy keletkeznek a

viccek? Miért van az, hogy bizonyos szavak viccesek, mások pedig nem? Hogy alakulnak ki a dialektusok? Le merem fogadni, hogy húsz év múlva már mélyreható ismereteink lesznek ezekről a témákról.

Azt hiszem azt is, fontolóra vehetjük, hogy a gyerekek online viselkedését – természetesen megfelelően anonimizálva – felhasználjuk a hagyományos vizsgálati módszerek kiegészítésére. Így könnyebben megfigyelhetjük, hogyan tanulnak és fejlődnek. Milyen a helyesírásuk? Nem mutatnak-e diszlexiára utaló jeleket? Érettebbek-e, alakul-e az intellektuális érdeklődésük? Vannak-e barátaik? Az ezekre a kérdésekre adható válaszokhoz azok a nyomok fognak elvezetni, amelyek ott rejlenek a minden egyes gyerek által minden egyes nap több ezer alkalommal leütött billentyűkben.

Van még egy további, kevésbé kézenfekvő terület, amelyen szintén rengeteg új felismerés várható.

A „Shattered” című Rolling Stones-dalban Mick Jagger elsorolja, mi minden teszi annyira varázslatossá New York Cityt, a Nagy Almát. Nevetés. Életöröm. Magány. Patkányok. Poloskák. Büszkeség. Kapzsiság. A papírzacskóba öltözött emberek. De Jagger a legtöbbször azt említi, ami igazán különlegessé teszi a várost: „sex and sex and sex and sex”.

Big Apple vagy Big Data – a lényeg ugyanaz. A digitális forradalom jóvoltából minden téren csőstül jönnek a felismerések. Alvás. Tanulás. Pszichológia. Nyelv... És a szex, a szex, a szex, a szex.

Íme, egy kérdés, amelyet éppen kutatók: hány dimenziója van a szexualitásnak? Általában úgy gondolunk adott személyre, mint heteróra vagy melegre. Csakhogy a szexualitás ennél sokkal komplexebb dolog. A meleg és a heterók csoportján belül is mindenkinek megvan a maga

esete – például egyiküknek a szőkék jönnek be, a másiknak pedig a barnák. Lehet, hogy ezek a preferenciák ugyanolyan erősek, mint a nemek szerinti preferencia? Egy másik kérdés, amellyel foglalkozom: honnan erednek a szexuális preferenciák? Mint kiderült, nemcsak a baseballszurkolóvá válás vagy a politikai nézetek, hanem a felnőttkori szexuális preferenciák esetében is meg tudjuk határozni, hogy melyek a rögzülés szempontjából kulcsfontosságú évek. Akit érdekelnek a most feltett kérdésekre adott válaszaim, annak meg kell vennie a következő könyvemet, amelynek munkacíme: *(Még mindig) mindenki hazudik*.

A pornó, illetve az azzal kapcsolatos adatok forradalmi fejleményt jelentenek az emberi szexualitás tudományos vizsgálatában. Időbe telt, mire a természettudományok lassanként megváltoztatták az életünket: létrehozták a penicillint, a műholdat, a számítógépet. Ugyanígy időbe telhet, mire a társadalomtudomány és a viselkedéstudomány a Big Data révén fontos felfedezéseket tesz arra vonatkozóan, hogy miként szeretünk, tanulunk, élünk. De hiszek benne, hogy a felfedezések már közel vannak. Szívből remélem, hogy a könyvem nyomán legalább a körvonalaiak felsejlenek önök előtt. Sőt, igazából azt is remélem, hogy olvasóim közül is kikerülnek majd olyanok, akik hozzátesznek valamit a dologhoz.

Aki rendes zárszót akar írni, annak el kell gondolkodnia azon, hogy miért is írta meg a könyvet. Mit akart vele elérni?

Azt hiszem, a könyvem elsősorban életem egyik sorsdöntő, személyiségformáló élményének köszönhető. Valamivel több mint egy évtizede történt, hogy megjelent a *Lökonómia* című könyv, amely mindenki meglepetésére bestseller lett. A *Lökonómia* az egyik társszerző, Steven

Levitt kutatásait ismerteti. A Chicagói Egyetem díjnyertes közgazdásza, akit én is sűrűn emlegetek ebben a könyvben, amolyan „elfajzott közgazdásznak” számított: arról híresült el, hogy szinte bármilyen kérdést képes megadotolni, ami csak kipattan abból a csavaros agyából. Csalnak-e a szumóbirkózók? Diszkriminálnak-e a tévévetélkedők játékosai? Ugyanazt nyújtják-e az ingatlanügynökök nekünk, mint saját maguknak?

Éppen akkoriban kerültem ki az egyetemről, ahol filozófia szakos voltam, és nemigen tudtam, mit akarok kezdeni az életemmel. De miután elolvastam a *Lökonómiát*, már tudtam: azt akartam csinálni, amit Steven Levitt. Én is adathegyeket akartam megmozgatni, hogy lássam, miként működik *valójában* a világ. Elhatároztam tehát, hogy Levitt nyomdokaiba lépek, és közgazdaságtanból fogok PhD-zni.

Az azóta eltelt tizenkét év alatt nagyot fordult a világ. Kiderült, hogy Levitt néhány tanulmányában kódolási hibákat követett el, és politikailag inkorrekt dolgokat mondott a globális felmelegedésről. A *Lökonómia* pedig intellektuális körökben kegyvesztett lett.

Ezzel együtt szerintem, hibák ide vagy oda, az eltelt évek kegyesek voltak a Levitt által kifejteni próbált álláspont-hoz. Levitt azt állította, hogy a kíváncsiság, a kreativitás és az adatok megfelelő kombinációja azt fogja eredményezni, hogy drámaian jobban fogjuk érteni a világot. Az adatokban elmesélésre váró sztorik rejlének, és ez újra meg újra beigazolódik.

Abban reménykedem, hogy talán az én könyvem is hathat másokra úgy, ahogy a *Lökonómia* hatott rám. Bízom benne, hogy akad valahol egy ifjú pályakezdő, aki ma még nem látja tisztán, mit is fog kezdeni az életével. Azt tudom tanácsolni neki, hogy ha konyít egy kicsit a statisztikához,

valamint nagy adag kreativitás és kíváncsiság szorult belé, tegyen egy próbát az adatelemző biznisszel.

A könyvem, ha lehetek ilyen merész, akár a *Lökonómia* következő szintjeként is felfogható. A *Lökonómiában*, illetve a *Mindenki hazudikban* ismertetett tanulmányok között leginkább az ambíció terén van különbség. hiszen a kilencvenes években, amikor Levitt megalapozta hírnevét, még nem volt túl sok hozzáférhető adat. Levitt büszke volt rá, hogy minden fogós kérdésnek a nyomába ered, ha talál róluk adatokat. Többé-kevésbé ignorálta azokat a kérdéseket, amelyekkel kapcsolatban nem voltak elérhető adatok. Ma azonban már szinte minden témáról sokkal, de sokkal több adat áll rendelkezésünkre, így van értelme, hogy célba vegyük az alapkérdéseket, és megpróbáljuk megérteni a maga teljes mélységében, mit is jelent embernek lenni.

Az adatelemzés fényes jövő előtt áll. Erősen gyanítom, hogy a következő Alfred Kinsey adattudós lesz. A következő Foucault adattudós lesz. A következő Freud adattudós lesz. A következő Marx adattudós lesz. Sőt, könnyen lehet, hogy a következő Jonas Salk is adattudós lesz.

Mindenesetre eddig futotta a próbálkozásomból, hogy valami olyasmit produkáljak, amit egy rendes zárszó szokott. Rá kellett jönnöm azonban, hogy igazán kiváló zárszónak ennél sokkal, de sokkal többet kell tudnia. Az igazán kiváló zárszó legyen ironikus. Legyen megható. Legyen precíz, alapos, ugyanakkor játékos. Legyen mélyenszántó, humoros és szomorú. Az igazán kiváló zárszónak egy-két velős, lényegre törő mondatba kell sűrítetnie a nagy összeggést, amely magába foglal mindent, ami addig volt, és mindent, ami még ezután következhet. Ebben a néhány mondatban lennie kell valami egyedi, újszerű ötletnek, afféle csavarnak. Egy kiváló könyv vége elmés, vicces, provokatív, és nagyot szól.

Talán most jött el az alkalmas pillanat, hogy szót ejtsék arról is, hogy néz ki nálam az írás folyamata. Nem vagyok valami bőbeszédű szerző. Ez a könyv is csupán hetvenöt-ezer szót tartalmaz, vagyis meglehetősen kurta egy ilyen gazdag témához képest.

A terjedelmi korlátaimat azonban megszállottsággal pótolom. A *The New York Times* szexrovatába írt első, kétezer szavas cikkemen öt hónapig dolgoztam, és ez idő alatt negyvenhét piszkozatot készítettem. Ennek a könyvnek van olyan fejezete, amely hatvan piszkozatot emésztett el. Képes vagyok órákig keresni a megfelelő szót egy lábjegyzet egyik mondatába.

Az elmúlt egy évben jórészt remeteéletet éltem. Csak én és a számítógépem. New York legmenőbb részén laktam, de jóformán ki sem mozdultam otthonról. Ez a könyv szerény véleményem szerint az én *magnum opusom*, ennél jobb ötletem aligha támad valaha is. Ezért kész voltam feláldozni mindent, amit csak kellett, hogy jól sikerüljön. Azt akartam, hogy minden egyes szavát meg tudjam védeni. A telefonom tele van megválaszolatlan e-mailekkel, megnyitatlan e-meghívókkal, olvasatlan Bumble-üzenetekkel.*

* Mivel mindenki hazudik, önök is nyugodtan kétségbe vonhatják a sztorim valóságtartalmát. Lehet, hogy nem is vagyok megszállott. Lehet, hogy nem is dolgoztam olyan rettentően keményen ezen a könyvön. Lehet, hogy én is, mint oly sokan, hajlamos vagyok eltúlozni, hogy mennyit dolgozom. Lehet, hogy a bő egy évi „kemény munkának” részét képezték teljes hónapok, amikor nem dolgoztam semmit. Lehet, hogy nem is éltem remeteéletet. Lehet, hogy ha valaki lecsékkolja a Facebook-profilomat, olyan fotókat fog látni, amelyeken a barátaimmal lógok az állítólagos remeteség idején. Vagy az is lehet, hogy remete voltam ugyan, de nem önszántamból. Lehet, hogy a magányos estéken sokszor képtelen voltam dolgozni, és epekedve vártam, hogy valaki keressen, de hiába. Lehet, hogy nem is küld nekem e-meghívót senki, és senki sem üzen nekem a Bumble-on. Mindenki hazudik. Minden elbeszélő megbízhatatlan.

Tizenhárom havi kemény munka után végre el tudtam küldeni egy majdnem kész piszkozatot. Egy fejezet azonban hiányzott: a zárszó.

Elmagyaráztam Denise-nek, a szerkesztőmnak, hogy ehhez szükség lehet még néhány hónapra. Azt mondtam neki, hogy ha muszáj tippelnem, úgy hatra. Elmondtam, hogy véleményem szerint a zárszó a könyv legfontosabb része, és én még csak most tanulgom, mi kell egy igazán kiváló konklúzióhoz. Talán mondanom sem kell, hogy Denise nem volt elragadtatva.

Aztán egy szép napon egy barátom elküldte nekem e-mailen Jordan Ellenberg tanulmányát. Ellenberg, a Wisconsin Egyetem matematikusa arra volt kíváncsi, hányan szoktak ténylegesen végigolvasni egy könyvet. Kieszelt egy ötletes módszert arra, hogy ezt Big Datával tesztelje. Azt használta ki, hogy az Amazon közli, hányan idézték a könyvek különböző mondatait. Rájött, hogy ha összehasonlítja a könyvek elején, illetve végén található idézetek gyakoriságát, azzal durván meg tudja becsülni, hogy az emberek mennyire hajlamosak kitartani a könyv végéig. Az elemzés szerint Donna Tartt *Az Aranyipinty* című regényét az olvasók több mint 90 százaléka végigolvasta. Ezzel szemben Daniel Kahneman, a Nobel-díjas közgazdász *magnum opusának*, a *Gyors és lassú gondolkodásnak* mindössze 7 százalék jutott a végére. Azok arányát pedig, akik végigolvasták egy másik közgazdász, Thomas Piketty *A tőke a 21. században* című, sokat vitatott és dicsért művét, Ellenberg hozzátételeges becslése 3 százalékra tette. Másként fogalmazva: nemigen jellemző, hogy az emberek átrágyák magukat a közgazdászok terjedelmes tanulmányain.

A könyvem egyik fontos üzenete az, hogy követnünk kell a Big Datát, *bárhová vezet minket*, és az útmutatásai

szerint kell eljárunk. Reménykedhetem benne, hogy legtöbb olvasóm minden szavamon csüggeni fog, és majd úgy igyekeznek mintázatokot felfedezni, hogy megpróbálnak kapcsolatot találni az utolsó oldalak és a korábbiak között. De csiszolgathatom én a prózai stílusomat napestig, a leg-
többen úgyis csak az első ötven oldalt olvassák el. Abból leszűrnek ezt-azt, majd élnek tovább az életüket.

Így hát az egyetlen ideillő módon fejezem be ezt a könyvet: az adatoknak engedelmeskedve, vagyis ahhoz tartva magam, amit az emberek valójában tesznek, és nem ahhoz, amit mondanak. Most elmegyek sörözni a barátaimmal, és abbahagyom ennek a nyomorult zárszónak az írását. A Big Data arról árulkodik, hogy önök közül úgyis csak nagyon kevesen jutottak el idáig az olvasásban.

Köszönetnyilvánítás

Ez a könyv csapatmunka eredménye.

Ezek a gondolatok harvardi diákéveim, a Google-nál való adattudóskodásom, valamint a *The New York Times*-nél való újságíróskodásom idején álltak össze.

Hal Varian, akivel együtt dolgoztunk a Google-nál, rendkívül nagy hatást gyakorolt a könyvben megfogalmazott gondolatokra. Amennyire meg tudom ítélni, Hal folyamatosan megelőzi a korát vagy húsz évvel. Az *információ uralma* című, Carl Shapiróval közösen írt könyvében gyakorlatilag megjósolta a jövőt, *Predicting the Present* (*Megjósolni a jelent*) című tanulmányával pedig, amelynek társszerzője Hyunyoung Choi, gyakorlatilag kirobbantotta a társadalomtudományok Big Data-forradalmát, amelyről a könyve szól. Hal mindezek mellett elképesztően jó és kedves mentor, ahogy azt oly sokan tanúsíthatják a vele dolgozók közül. Klasszikus Hal-módi, hogy oroszlánrészt vállal a tanulmány elkészítésében, amelynek te vagy a társszerzője, majd erősködik, hogy ne az ő neve álljon elől, hanem a tiéd. Hal a zsenialitás és a nagylelkűség olyasféle kombinációja, amelyhez foghatóval nemigen találkoztam.

Az íráskészségem és a gondolatvilágom is sokat fejlődött Aaron Retica szárnyai alatt, aki minden egyes *The New York Times*-cikkem szerkesztője volt. Aaron igazi polihisztor. Valahogy mindent tud a zenéről, a történelemtől, a

sportról, a politikáról, a szociológiáról, a közgazdaságtanról, meg még a jó ég tudja, mi mindenről. Nagyrészt ő a felelős mindazért, ami a nevem alatt megjelent *Times*-cikkekben jónak mondható. A rovat csapatában játszott még Bill Marsh is, akinek grafikai megoldásaitól mindmáig lehidalogok, valamint Kevin McCarthy és Gita Daneshjoo. Ez a könyv több, a lap engedélyével átvett részletet tartalmaz ezekből a cikkekből.

Steven Pinker, aki volt szíves megírni az előszót, régi bálványom. Ő szabta meg a mintát a modern társadalomtudományi könyvek számára: az emberi természet alapjainak sodró erejű feltárása, számos tudományág kutatásainak legjavát értelmezve a felfedezés során. Egész életemben azért fogok küzdeni, hogy felnőhessek ehhez a mintához.

A disszertációm, amelyből ez a könyv kinőtt, zseniális és türelmes konzulenseim, Alberto Alesina, David Cutler, Ed Glaeser és Lawrence Katz segítségével írtam.

Denise Oswald bámulatosan jó szerkesztő. Ha tudni akarják, mennyire jól dolgozik, hasonlítsák össze a végleges kéziratot az első verzióval – na jó, ez igazából nem áll módjukban, mert azt a kínos első piszkozatot soha, senkinek nem fogom megmutatni. Köszönettel tartozom a HarperCollins-csapat többi tagjának is, köztük Michael Barrsnek, Lynn Gradynek, Lauren Janiecnek, Shelby Meizliknek és Amber Olivernek is.

Eric Lupfer, az ügynököm kezdettől fogva látott fantáziát a projektben. Közreműködött a kiadónak beadott könyvterv kidolgozásában, és a megvalósításban is végig a segítségemre volt.

Az egészen kiváló tényellenőrzésért Melvis Acostát illeti köszönet.

Szakmai és tudományos pályám során sokat tanultam mások mellett az alábbiaktól: Susan Athey, Shlomo Be-

nartzi, Jason Bordoff, Danielle Bowers, David Broockman, Bo Cowgill, Steven Delpome, John Donohue, Bill Gale, Claudia Goldin, Suzanne Greenberg, Shane Greenstein, Steve Grove, Mike Hoyt, David Laibson, A.J. Magnuson, Dana Maloney, Jeffrey Oldham, Peter Orszag, David Reiley, Jonathan Rosenberg, Michael Schwarz, Steve Scott, Rich Shavelson, Michael D. Smith, Lawrence Summers, Jon Vaver, Michael Wiggins és Qing Wu.

Az íráskészségem fejlesztésében nyújtott segítségért Tim Requarthnak és a NeuWrite-nak tartozom hálával.

A tanulmányok értelmezésében nyújtott segítségükért Christopher Chabrisnek, Raj Chettynek, Matt Gentzkow-nak, Solomon Messingnek és Jesse Shapirónak mondok köszönetet.

Emma Piersontól és Katia Sobolskitól megkérdeztem, tudnának-e esetleg tanácsot adni az egyik fejezettel kapcsolatban. Erre ők, számomra érthetetlen okokból, felajánlották, hogy elolvassák az egészet – és minden egyes bekezdésre vonatkozóan bölcs tanácsokkal szolgáltak.

Édesanyám, Esther Davidowitz több alkalommal is végigolvasta az egész könyvet, és segített lényegesen jobbra alakítani. Anyukám tanított meg például arra, hogy bárhová visz is, mindig a kíváncsiságom vezéreljen. Egy egyetemi állásinterjún a professzor nekem szegezte a kérdést: „Az édesanyja mit szól a munkájához?” Az volt az elképzelése, hogy majd jól zavarba jövök, hiszen anyukámat nyilván kínosan érinti, hogy szexszel és más tabutémákkal foglalkozom. De én mindig is tudtam, hogy anyukám büszke rám, amiért bárhová visz is, mindig a kíváncsiságom vezérel.

Sokan olvasták el a könyv egyes részeit és fűztek hozzá hasznos kommentárokat. Köszönettel tartozom Eduardo Acevednek, Coren Apicellának, Sam Ashernek, David

Cutlernek, Stephen Dubnernek, Christopher Glazeknek, Jessica Goldbergnek, Lauren Goldmannek, Amanda Gordonnak, Jacob Leshnónak, Alex Peysakhovichnak, Noah Poppnak, Ramon Roullardnak, Greg Sobolskinak, Evan Soltasnak, Noah Stephens-Davidowitznak, Lauren Stephens-Davidowitznak, valamint Jean Yangnek, aki gyakorlatilag a legjobb barátom volt a könyvírás időszakában, így ezt is megköszönöm neki.

Az adatgyűjtésben nyújtott segítségért köszönettel tartozom Brett Goldenbergnek, James Rogersnek és Mike Williamsnek (MindGeek), valamint Rob McQuownnak és Sam Millernek (Baseball Prospectus).

Hálásan köszönöm az Alfred Sloan Alapítvány anyagi támogatását.

A könyv írása során egy ponton nagyon elakadtam; úgy éreztem, tévúton járok, és közel jártam ahhoz, hogy az egészet félbehagyjam. És akkor apámmal, Mitchell Stephensszel elmentünk vidékre, ahol apu egy hét leforgása alatt helyrerakott. Hosszú sétákat tettünk, és nagyokat beszélgettünk szerelemről, halálról, sikerről, boldogságról és írásról – aztán pedig apu leültetett, és mondatról mondatra végigmentünk a könyvön. Nélküle nem tudtam volna befejezni.

Természetesen minden, a könyvben megmaradt hibáért engem terhel a felelősség.

Jegyzetek

BEVEZETÉS

12. oldal az amerikai szavazókat nagyrészt hidegen hagyta, hogy Barack Obama: Katie Fretland: „Gallup: Race Not Important to Voters”, *The Swamp, Chicago Tribune*, 2008. június.
12. oldal Berkeley [...] ásta bele magát: Alexandre Mas és Enrico Moretti: „Racial Bias in the 2008 Presidential Election”, *American Economic Review* 99, no. 2 (2009).
12. oldal posztrasszista társadalomban: Lou Dobbs a műsora 2009. november 12-i adásában közölte: „posztfanatikus, posztrasszista társadalomban” élünk. Chris Matthews a műsora 2010. január 27-i adásában közölte, hogy Obama elnök „minden jel szerint posztrasszista”. További példákat ld. Michael C. Dawson és Lawrence D. Bobo: „One Year Later and the Myth of a Post-Racial Society”, *Du Bois Review: Social Science Research on Race* 6, no. 2 (2009).
15. oldal Elemeztem [...] a General Social Survey adatait: A számítások részletes ismertetése megtalálható a honlapomon (sethsd.com), a „Sex Data” címkét viselő táblázatban. A General Social Survey adatai elérhetők a <http://gss.norc.oregon.edu/webhelyen>.
16. oldal nem éri el a 600 milliót: A szerző rendelkezésére bocsátott adat.
17. oldal Stormfront [...] kerestek rá, illetve iratkoztak fel: A Google Trends-adatoknak a szerző által elvégzett elemzése. Összekapartam továbbá a Stormfront összes tagjának adatait, amint azt az alábbi cikkemben kifejtem: Seth Stephens-Davidowitz: „The Data of Hate”, *The New York Times*, 2014. július 13., SR 4. A vonatkozó adatok letölthetők a sethsd.com oldalon, a „Stormfront” címkéjű adatmellékletben.

18. oldal többször kerestek rá arra, hogy „nigger elnök”, mint arra, hogy „első fekete elnök”: A Google Trends-adatoknak a szerző által elvégzett elemzése. Az állítás többek között Kentucky, Louisiana, Arizona és Észak-Karolina államokra igaz.
20. oldal öt tudományos folyóirat is elutasította: A tanulmány végül Seth Stephens-Davidowitz: „The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data” címen jelent meg a *Journal of Public Economics* 118. (2014)-es számában. A kutatás további részleteit ld. ott. Az adatok megtalálhatók továbbá a honlapomon (sethsd.com), a „Racism” című adatmellékletben is.
24. oldal az az egy tényező, amely a leginkább korrelált: „A legerősebb korreláció, amit Trump támogatottságára találtam, az »n« betűs szóra irányuló Google-keresések voltak. Erről mások is beszámoltak” (Twitter-poszt, 2016. február 28.). Ld. még Nate Cohn: „Donald Trump’s Strongest Supporters: A New Kind of Democrat”, *The New York Times*, 2015. december 31., A3.
25. oldal A diagram a „nigger(ek)” szót tartalmazó Google-keresések százalékos arányát mutatja. Fontos, hogy mivel százalékos arányról van szó, a szám nem magasabb a népesebb területeken vagy ott, ahol sok keresést indítanak az emberek. Fontos továbbá, hogy a Trump támogatottsági térképéhez képest mutatkozó különbségre kézenfekvő a magyarázat: Trump azért volt kevésbé népszerű Texasban és Arkansasban, mert ebből a két államból származott két ellenfele, Ted Cruz és Mike Huckabee.
25. oldal Az adatok a Civis Analytics 2015. decemberi felméréséből származnak. A tényleges szavazási adatok kevésbé hasznosak, mivel azokat erősen befolyásolja az előválasztás időpontja, illetve a szavazás módja. A térképek újraközlése a *The New York Times* engedélyével történt.
27. oldal 2,5 millió trillió bájtnyi adatot: „Bringing Big Data to the Enterprise”, IBM, <https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>.
29. oldal egyre nagyobb szénakazalban kell keresnünk a tűt: Nassim M. Taleb: „Beware the Big Errors of ‘Big Data’”, *Wired*, 2013. február 8., <http://www.wired.com/2013/02/big-data-means-big-errors-people>.
30. oldal sem a rasszista keresések száma, sem a Stormfront taglét-

száma: Megvizsgáltam, miként változott az internetes rasszizmus azokban az országrészekben, amelyeket jobban, illetve kevésbé sújtott a válság. Megnéztem még a „nigger(ek)” keresőszó előfordulási arányát, valamint a Stormfront-tagságot is. A vonatkozó adatok letölthetők a sethsd.com oldalon, a „Racial Animus” és a „Stormfront” című adatmellékletekben.

30. oldal szorongással kapcsolatos Google-keresések: Seth Stephens-Davidowitz: „Fifty States of Anxiety”, *The New York Times*, 2016. augusztus 7., SR 2. Fontos, hogy a Google-keresések sokkal nagyobb mintát szolgáltatnak ugyan, de ez a mintázat konzisztens a felmérések adataival. Ld. pl. William C. Reeves et al.: „Mental Illness Surveillance Among Adults in the United States”, *Morbidity and Mortality Weekly Report Supplement* 60, no. 3 (2011).
31. oldal olyankor keresnek gyakrabban vicceket: Ezt az alábbi cikkemben tárgyalom: Seth Stephens-Davidowitz: „Why Are You Laughing?”, *The New York Times*, 2016. május 15. SR 9. A vonatkozó adatok letölthetők a sethsd.com oldalon, a „Jokes” című adatmellékletben.
31. oldal „a férjem azt akarja, hogy szoptassam”: Ezt az alábbi cikkemben tárgyalom: Seth Stephens-Davidowitz: „What Do Pregnant Women Want?”, *The New York Times*, 2014. május 17., SR 6.
31. oldal a férfit szoptató nőt ábrázoló pornóra való keresés: A Pornhub-adatoknak a szerző által elvégzett elemzése.
32. oldal a nőknél majdnem ugyanannyi: Ezt az alábbi cikkemben tárgyalom: Seth Stephens-Davidowitz: „Searching for Sex”, *The New York Times*, 2015. január 25., SR 1.
32. oldal „poemas para mi esposa embarazada”: Stephens-Davidowitz: „What Do Pregnant Women Want?”.
33. oldal mondja Friedman: 2015. október 27-én telefoninterjút készítettem Jerry Friedmannal.
34. oldal összes adatból vett parányi minta: Hal R. Varian: „Big Data: New Tricks for Econometrics”, *Journal of Economic Perspectives* 28, no. 2 (2014).

1. FEJEZET: ÖSZTÖNÖS, FÉLREÉRZÉSEINK

40. oldal Sőt, ami azt illeti, az adattudomány meglepően intuitív tevékenység: az adatelemzésnek arról a szegletéről beszélék,

- amelyiket ismerem: az emberi viselkedést magyarázni és előre jelezni próbáló adattudományról. Azaz nem a mesterséges intelligenciáról, amely megpróbál, mondjuk, autót vezetni. Az efféle módszerek, bár alkalmaznak olyan eszközöket, amelyeket az emberi agy működéséből kiindulva találtak fel, nem ilyen könnyen érthetők.
42. oldal *milyen tünetek jelzik előre a hasnyálmirigyrákot*: John Parrizos, Ryan W. White és Eric Horvitz: „Screening for Pancreatic Adenocarcinoma Using Signals from Web Search Logs: Feasibility Study and Results”, *Journal of Oncology Practice* (2016).
46. oldal *A téli időjárás hatása minden más tényezőét kenterbe verte*: Ezt a kutatást az alábbi cikkemben tárgyalom: Seth Stephens-Davidowitz: „Dr. Google Will See You Now”, *The New York Times*, 2013. augusztus 11., SR 12.
47. oldal *az emberi kapcsolatokról valaha felhalmozott legnagyobb adathalmazt*: Lars Backstrom és Jon Kleinberg: „Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook”, in: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2014).
48. oldal *az emberek következetesen előbbre helyezik*: Daniel Kahneman: *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux, 2011. (Magyarul: *A gyors és lassú gondolkodás*, HVG Könyvek, 2012.)
48. oldal *az asztma valójában hetvenszer több ember életét oltja ki*: az Egyesült Államokban 1979 és 2010 között évente átlagosan 55,81 volt a tornádó miatt elhalálozottak száma, míg 4216,53 az asztma miatt elhalálozottak száma. Ld. az Amerikai Meteorológiai Szolgálat éves tornádóhalál-statisztikáját (<http://www.spc.noaa.gov/climo/torn/fatalmap.php>), valamint az Amerikai Tüdőgyógyász Társaság epidemiológiai és statisztikai csoportjának Trends in Asthma Morbidity and Mortality című anyagát.
48. oldal *Patrick Ewing*: A kedvenc Ewing-videóim a következők: „Patrick Ewing’s Top 10 Career Plays”, YouTube-videó, posztolva 2015. szeptember 18-án, <https://www.youtube.com/watch?v=Y-29gMuYymv8>; valamint „Patrick Ewing Knicks Tribute”, YouTube-videó, posztolva 2006. május 12-én (<https://www.youtube.com/watch?v=8T2l5Emzu-I>).

49. oldal *a kosárlabda élet-halál kérdése*: S. L. Price: „Whatever Happened to the White Athlete?”, *Sports Illustrated*, 1997. december 8.
49. oldal *internetes felmérésből*: A Google Consumer Surveys segítségével, 2013. október 22-én végzett felmérésről van szó. Ezt a kérdést tettem fel: „Mit tippel, hol született az NBA-kosarasok többsége?” A megkérdezettek kétféle választ adhattak: „szegények lakta környéken”, illetve „középosztály lakta környéken”. 59,7 százalékuk választotta a szegények lakta környéket.
51. oldal *a feketék keresztneve jelzésértékű a szocioökonómiai háttérük szempontjából*: Roland G. Fryer Jr. és Steven D. Levitt: „The Causes and Consequences of Distinctively Black Names”, *Quarterly Journal of Economics* 119, no. 3 (2004).
52. oldal *a nyolcvanas években született afroamerikaiak*: Járványvédelmi és -Megelőzési Központ: „Health, United States, 2009”, 9. táblázat: Házasságon kívüli gyerekszülési adatok az anya faji hovatartozása, illetve hispanikus származása, valamint életkora szerint: Egyesült Államok, kiválasztott évek: 1970–2006.
53. oldal *Chris Bosh [...] Chris Paul*: „Not Just a Typical Jock: Miami Heat Forward Chris Bosh’s Interests Go Well Beyond Basketball”, *Palm BeachPost.com*, 2011. február 15, <http://www.palm-beachpost.com/news/sports/basketball/not-just-a-typical-jock-miami-heat-forward-chris-b/nLp7Zl/>; Dave Walker: „Chris Paul’s Family to Compete on ‘Family Feud’”, *nola.com*, 2011. október 31., http://www.nola.com/tv/index.ssf/2011/10/chris-pauls_family_to_compete.html.
53. oldal *tíz centivel magasabb*: „Why Are We Getting Taller as a Species?”, *Scientific American*, <http://www.scientificamerican.com/article/why-are-we-getting-taller/>. Csakhogy az amerikaiak érdekes módon megálltak a növésben, ld. Amanda Onion: „Why Have Americans Stopped Growing Taller?”, *ABC News*, 2016. július 3., <http://abcnews.go.com/Technology/story?id=98438&page=1>. Én amellett érveltem, hogy a külföldről származó NBA-játékosok száma részben azért nőtt meg óriási mértékben, mert más országok is kezdenek felzárkózni az Egyesült Államokhoz a testmagasság terén. A 213 centinél magasabb, amerikai születésű játékosok száma 16-szorosára nőtt 1946 és 1980 között, amikor az amerikaiak egyre csak nőttek.

- Azóta ez a szint rögzült, mivel az amerikaiak megálltak a növekedésben. Mindeközben a 213 centi feletti külföldiek száma jelentős mértékben növekedett. Mint kiderült, ez főként az olyan országokból (pl. Törökország, Spanyolország, Görögország) érkezett kosarasoknak köszönhető, ahol a megelőző években észrevehető javulás állt be a gyerekkori egészség, valamint a felnőttkori testmagasság terén.
53. oldal *szegény sorsú amerikaiak*: Carmen R. Isasi et al.: „Association of Childhood Economic Hardship with Adult Height and Adult Adiposity among Hispanics/Latinos: The HCHS/SOL Socio-Cultural Ancillary Study”, *PLoS One* 11, no. 2 (2016); Jane E. Miller és Sanders Korenman: „Poverty and Children’s Nutritional Status in the United States”, *American Journal of Epidemiology* 140, no. 3 (1994); Harry J. Holzer, Diane Whitmore Schanzenbach, Greg J. Duncan és Jens Ludwig: „The Economic Costs of Childhood Poverty in the United States”, *Journal of Children and Poverty* 14, no. 1 (2008).
53. oldal *az átlagos amerikai férfi 5 láb 9 hüvelyk*: Cheryl D. Fryar, Qiuping Gu és Cynthia L. Ogden: „Anthropometric Reference Data for Children and Adults: United States, 2007–2010”, *Vital and Health Statistics Series* 11, no. 252 (2012).
54. oldal *nagyjából minden ötödik*: Pablo S. Torre: „Larger Than Real Life”, *Sports Illustrated*, 2011. július 4.
55. oldal *kétszülős, középosztálybeli családokból*: Tim Kautz, James J. Heckman, Ron Diris, Bas Ter Weel és Lex Borghans: „Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success”, a Nemzeti Gazdaságkutató Hivatal 20749. számú munkaanyaga, 2014.
55. oldal *Wrenn magasabbra ugrik, mint bárki*: Desmond Conner: „For Wrenn, Sky’s the Limit”, *Hartford Courant*, 1999. október 21.
55. oldal *Csakhogy Wrenn*: Doug Wrenn történetét ld. Percy Allen: „Former Washington and O’Dea Star Doug Wrenn Finds Tough Times”, *Seattle Times*, 2009. március 29.
55. oldal „Doug Wrenn halott”: uo.
56. oldal *Jordan is tudott nehéz eset lenni*: Melissa Isaacson: „Portrait of a Legend”, ESPN.com, 2009. szeptember 9., http://www.espn.com/chicago/columns/story?id=4457017&columnist=isaacson_

melissa. Egy jó Jordan-életrajz: Roland Lazenby: *Michael Jordan: The Life* (Boston, Back Bay Books, 2015).

56. oldal *Az apja*: Barry Jacobs: „High-Flying Michael Jordan Has North Carolina Cruising Toward Another NCAA Title”, *People*, 1984. március 19.
56. oldal *Jordan élettörténete tele van olyan sztorikkal, amelyek arról szólnak, miként óvta őt a családja*: Isaacson: *Portrait of a Legend*.
56. oldal *a hírességek Csarnokában szólalt fel a bevélasztása alkalmából*: Michael Jordan’s Basketball Hall of Fame Enshrinement Speech, YouTube-videó, posztolva 2012. február 21-én, <https://www.youtube.com/watch?v=XLzBMGXfK4c>. A beszéd legérdekesebb vonása nem is az, hogy Jordan terjengősen áradozik a szüleiről: inkább az, hogy még ekkor is szükségesnek tartotta felemlgetni a pályája kezdetén elszenvedett sérelmeket. Talán a sérelmek élethosszig tartó dédelgetése előfeltétele annak, hogy valaki minden idők legnagyobb kosarasává váljon.
57. oldal *LeBron James interjút adott*: „I’m LeBron James from Akron, Ohio”, YouTube-videó, posztolva 2013. június 20-án, <https://www.youtube.com/watch?v=XceMbPVAggk>.

2. FEJEZET: IGAZA VOLT-E FREUDNAK?

62. oldal *fallosz formájú étel*: Azokat az ennivalókat kódoltam fallosz formájuként, amelyek lényegesen hosszabbak, mint amilyen szélesek, és alapvetően kereksek. Az uborka, a kukorica, a sárgarépa, a padlizsán, a süttők és a banán előfordulásait számoltam. Az adatok, a kódolás és a kutatás részletes leírása megtalálható a sethsd.com oldalon.
64. oldal *a Microsoft kutatói által összeállított [...] gépelési hibát*: Az adathalmaz letölthető innen: <https://www.microsoft.com/en-us/download/details.aspx?id=52418>. A kutatók megkérték az Amazon Mechanical Turk felhasználóit, hogy írjanak le különböző képeket. Ezután elemezték a billentyűleütés-naplókat, és rögzítették a javításokat. További részleteket ld. Yukino Baba és Hisami Suzuki: „How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs”, a Számítógépes Nyelvészeti Társaság 50. éves

gyűlésének jegyzőkönyve, 2012. Az adatok, a kódolás, valamint a kutatás részletes leírása megtalálható a sethsd.com oldalon.
67. oldal „szexelni akarok a...” kezdetű kereséseket: A teljes adathalmaz – vigyázat, szókimondó tartalom! – így néz ki:

„SZEKELNI AKAROK...”

	GOOGLE-KERESÉSEK HAVONTA, ERRE A KONKRÉT KIFEJEZÉSRE
az anyámmal	720
a fiammal	590
a lánytestvéremmel	590
az unokatestvéremmel	480
az apámmal	480
a pasimmal	480
a fiútestvéremmel	320
a lányommal	260
a barátommal	170
a barátnőmmel	140

68. oldal *rajzfilmpornó*: Példának okáért, amint az az alábbi táblázatból is kiténik, a „pornó” az egyik leggyakrabban előforduló szó a különböző, különösen nagy népszerűségnek örvendő animációs sorozatokra irányuló Google-keresésekben.

RAJZFILMEK, ISMERKEDJETEK MEG A PORNÓVAL!
(A KÜLÖNBÖZŐ RAJZFILMEKRE IRÁNYULÓ
LEGGYAKORIBB GOOGLE-KERESŐKIFEJEZÉSEK)

family guy pornó	simpson családot nézni	futurama pornó	scooby doo játékok
family guy epizódok	simpson család pornó	futurama leela	scooby doo film
family guy ingyen	simpson család online	futurama epizódok	scooby doo pornó
family guyt nézni	simpson család film	futurama online	scooby doo vilma

69. oldal *bébiszitterként*: A szerző számításai szerint a férfiak által indított pornókeresésekben az alábbi foglalkozásokat űző nők a legnépszerűbbek (a férfiak életkora szerinti bontásban):

A NŐK FOGLALKOZÁSA A FÉRFIAK ÁLTAL INDÍTOTT
PORNÓKERESÉSEKBEN, A FÉRFIAK ÉLETKORA SZERINT

	18–24	25–64	65+
1.	Bébiszitter	Bébiszitter	Bébiszitter
2.	Tanárnő	Jógaoktató	Pomponlány
3.	Jógaoktató	Tanárnő	Doktornő
4.	Pomponlány	Pomponlány	Tanárnő
5.	Doktornő	Ingatlanügynök	Ingatlanügynök
6.	Prostituált	Doktornő	Nővér
7.	Ingatlanügynök	Prostituált	Jógaoktató
8.	Nővér	Titkárnő	Titkárnő
9.	Titkárnő	Nővér	Prostituált

3. FEJEZET: ADATOK ÚJRAGONDOLVA

73. oldal *cégek algoritmusai*: Matthew Leising: „HFT Treasury Trading Hurts Market When News Is Released”, Bloomberg Markets, 2014. december 16.; Nathaniel Popper: „The Robots Are Coming for Wall Street”, *The New York Times Magazine*, 2016. február 28., MM56. o; Richard Finger: „High Frequency Trading: Is It a Dark Force Against Ordinary Human Traders and Investors?”, *Forbes*, 2013. szeptember 30, <http://www.forbes.com/sites/richardfinger/2013/09/30/high-frequency-trading-is-it-a-dark-force-against-ordinary-human-traders-and-investors/#50875fc751a6>.

73. oldal *Alan Krueger*: Alan Kruegerrel telefonon készítettem interjút 2015. május 8-án.

74. oldal *fontos indikátornak bizonyultak a kór terjedési sebességére vonatkozóan*: Az eredeti tanulmány a következő volt: Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski és Larry Brilliant: „Detecting Influenza Epidemics Using Search Engine Query Data”, *Nature* 457, no.

- 7232 (2009). Az eredeti modell fogyatékoságait illetően ld. David Lazer, Ryan Kennedy, Gary King és Alessandro Vespignani: „The Parable of Google Flu: Traps in Big Data Analysis”, *Science* 343, no. 6176 (2014). A javított modellt mutatja be a következő cikk: Shihao Yang, Mauricio Santillana és S. C. Kou: „Accurate Estimation of Influenza Epidemics Using Google Search Data Via ARGO”, *Proceedings of the National Academy of Sciences* 112, no. 47 (2015).
75. oldal mely keresések követik a legjobban a lakásárak alakulását: Seth Stephens-Davidowitz és Hal Varian: „A Hands-on Guide to Google Data, stencilezett anyag, 2015. Ld. még Marcelle Chauvet, Stuart Gabriel és Chandler Lutz: „Mortgage Default Risk: New Evidence from Internet Search Queries”, *Journal of Urban Economics* 96 (2016).
77. oldal Bill Clinton: Sergey Brin és Larry Page: „The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Seventh International World-Wide Web Conference, 1998. április 14–18., Brisbane, Ausztrália.
78. oldal pornóoldalak: John Battelle: *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, New York, Penguin, 2005. (Magyarul: *Keress! Hogyan alakítja át kultúránkat, üzleti életünket a Google és az internetes keresés*. HVG Könyvek, 2006.)
79. oldal kiszervezte a rangsorolást: Jól tárgyalja a témát a következő könyv: Steven Levy: *In the Plex: How Google Thinks, Works, and Shapes Our Lives*, New York, Simon & Schuster, 2011.
81. oldal „Inkább a házát adja el”: Az idézet szerepel még Joe Drape: „Ahmed Zayat’s Journey: Bankruptcy and Big Bets” című cikkében is, *The New York Times*, 2015. június 5., A1. A cikk azonban tévesen Jeff Sedernek tulajdonítja ezt a szöveget, pedig valójában az egyik munkatársa mondta.
82. oldal találkoztam először (...) Sederrel: Jeff Sederrel és Patty Murray-vel a floridai Ocalában készítettem interjút 2015. június 12. és 14. között.
83. oldal Durván egyharmadukról: a versenylovak kudarcának okait Seder a szakmában eltöltött hosszú évekre épülő, durva becslésre alapozta.
83. oldal évente több száz ló pusztul el: Supplemental Tables of Equine Injury Database Statistics for Thoroughbreds, http://jockeyclub.com/pdfs/eid_7_year_tables.pdf.
83. oldal javarészt lábtörés miatt: „Postmortem Examination Program”, Kaliforniai Állategészségügyi és Élelmiszer-laboratóriumi Rendszer, 2013.
84. oldal több mint háromnegyedük soha életében nem nyer komolyabb versenyt: Avalyn Hunter: „A Case for Full Siblings”, *Bloodhorse*, 2014. április 18., <http://www.bloodhorse.com/horse-racing/articles/115014/a-case-for-full-siblings>.
85. oldal Earvin Johnson III-at: Melody Chiu: „E. J. Johnson Loses 50 Lbs. Since Undergoing Gastric Sleeve Surgery”, *People*, 2014. október 1.
85. oldal LeBron James, akinek az anyukája 165,1 centi magas: Eli Saslow: „Lost Stories of LeBron, Part 1”, ESPN.com, 2013. október 17., http://www.espn.com/nba/story/_/id/9825052/how-lebron-james-life-changed-fourth-grade-espn-magazine.
86. oldal The Green Monkey: Ld. Sherry Ross: „16 Million Dollar Baby”, *New York Daily News*, 2006. március 12. és Jay Privman: The Green Monkey, Who Sold for \$16M, Retired”, ESPN.com, 2008. február 12., <http://www.espn.com/sports/horse/news/story?id=3242341>. Az árverésről készült videó elérhetősége: „\$16 Million Horse”, YouTube-videó, posztolva 2008. november 1-én, <https://www.youtube.com/watch?v=EyggMC85Zsg>.
89. oldal A Google törekvésének, hogy a keresési adatok segítségével előre jelezze az influenzát, az volt a gyenge pontja: Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock és Duncan J. Watts: „Predicting Consumer Behavior with Web Search”, *Proceedings of the National Academy of Sciences* 107, no. 41 (2010).
90. oldal Az epres Pop-Tarts töltött keksz: Constance L. Hays: „What Wal-Mart Knows About Customers’ Habits”, *The New York Times*, 2004. november 14.
92. oldal „remekül bevált a dolog”: Orley Ashenfelterrel telefonon készítettem interjút 2016. október 27-én.
99. oldal több száz heteroszexuális villámrandizót vizsgáltak: Daniel A. McFarland, Dan Jurafsky és Craig Rawlings: „Making the

101. oldal *Leonard Cohen* Cohen [...] az unokaöccsének például ezt a tanácsot adta az udvarlással kapcsolatban: Jonathan Greenberg: „What I Learned From My Wise Uncle Leonard Cohen”, *Huffington Post*, 2016. november 11.

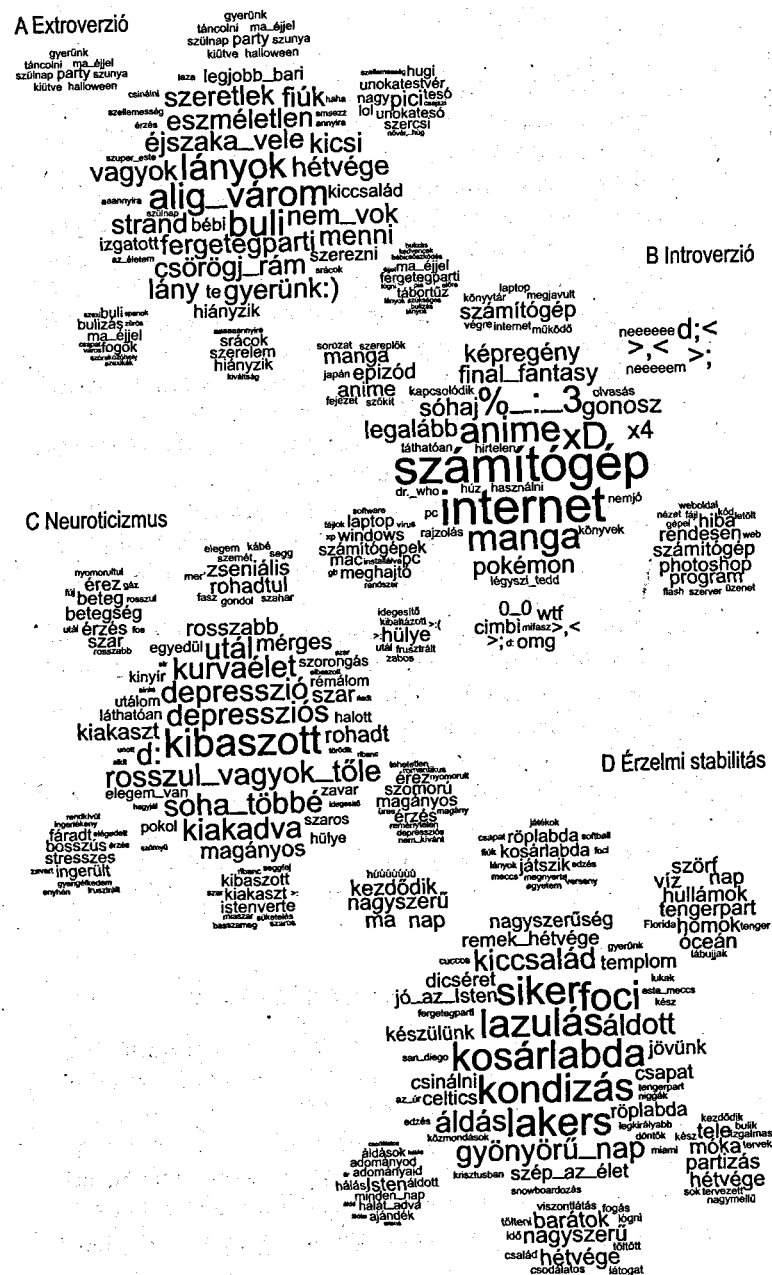
102. oldal *több százezer Facebook-poszt szavait elemezte*: H. Andrew Schwartz et al.: „Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach”, *PloS One* 8, no. 9 (2013). A tanulmány a személyiségteszten elért pontszámok szerint is osztályozza az emberek beszédmódját. Íme, az eredmény (szemben).

108. oldal *több ezer könyv és filmforgatókönyv szövegét*: Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth és Peter Sheridan Dodds: „The Emotional Arcs of Stories Are Dominated by Six Basic Shapes”, *EPJ Data Science* 5, no. 1 (2016).

111. oldal *milyen típusú történeteket osztanak meg*: Jonah Berger és Katherine L. Milkman: „What Makes Online Content Viral?”, *Journal of Marketing Research* 49, no. 2 (2012).

116. oldal *miért húznak egyes kiadványok balra*: A kutatás részletes kifejtését ld. Matthew Gentzkow és Jesse M. Shapiro: „What Drives Media Slant? Evidence from U. S. Daily Newspapers”, *Econometrica* 78, no. 1 (2010). Gentzkow és Shapiro ma már sztárközgazdászok, de a projekt kezdetén még csak PhD-hallgatók voltak. Gentzkow, aki ma a Stanford professzora, 2014-ben elnyerte a legjobb negyven év alatti közgazdásznak járó John Bates Clark-érmet. Shapiro pedig a Brown Egyetem professzora, valamint a *Journal of Political Economy* című tekintélyes folyóirat szerkesztője. A média elfogultságáról írt közös tanulmányuk mindkettőjük legidézettebb cikkei közé tartozik.

116. oldal *Rupert Murdoch*: Az, hogy a konzervatív *New York Post*-nak Murdoch a tulajdonosa, azzal magyarázható, hogy New York hatalmas metropolisz, így több, különböző irányultságú lapot is képes eltartani. Elég egyértelműen látszik azonban, hogy a *Post* folyamatosan veszteséges. Ld. pl. Joe Pompeo: „How Much Does the ‘New York Post’ Actually Lose?”, *Politico*, 2013. augusztus 30., <http://www.politico.com/media /story/2013/08/how-much-does-the-new-york-post-actually-lose-001176>.



118. oldal *mondta nekem Shapiro*: Matt Gentzkow-val és Jesse Shapiro-val a bostoni Royal Sonesta Hotelben készítettem interjút 2015. augusztus 16-án.
119. oldal *számos középiskola beszkenelte (...) az évkönyveit*: Kate Rakelly, Sarah Sachs, Brian Yin és Alexei A. Efros: „A Century of Portraits: A Visual Historical Record of American High School Yearbooks”. A tanulmányt 2015-ben mutatták be a számítógépes látásról rendezett nemzetközi konferencián. A fotók újraközlése a szerzők engedélyével történt.
120. oldal *a fényképek modelljei a festmények modelljeit utánozták*: Ld. pl. Christina Kotchemidova: „Why We Say ‘Cheese’: Producing the Smile in Snapshot Photography”, *Critical Studies in Media Communication* 22, no. 1 (2005).
121. oldal *az éjszakai fényerősség alapján fogják mérni a GDP-t*: J. Vernon Henderson, Adam Storeygard és David N. Weil: „Measuring Economic Growth from Outer Space”, *American Economic Review* 102, no. 2 (2012).
122. oldal *a becsült GDP így 90 százalékkal magasabb lett*: Kathleen Caulderwood: „Nigerian GDP Jumps 89% as Economists Add in Telecoms, Nollywood”, *IBTimes*, április 7., 2014, <http://www.ibtimes.com/nigerian-gdp-jumps-89-economists-add-telecoms-nollywood-1568219>.
122. oldal *mondta lassan emelkedő hangerővel Reisinger*: Joe Reisingerrel telefonon készítettem interjút 2015. június 10-én.
124. oldal *50 millió dollár*: Leena Rao: „SpaceX and Tesla Backer Just Invested \$50 Million in This Startup”, *Fortune*, 2015. szeptember 24.

4. FEJEZET: DIGITÁLIS IGAZSÁGSZÉRUM

126. oldal *Már 1950-ben készült egy fontos tanulmány*: Hugh J. Parry és Helen M. Crossley: „Validity of Responses to Survey Questions”, *Public Opinion Quarterly* 14, 1 (1950).
127. oldal *a Marylandi Egyetemen végzett felmérés keretében számos kérdést feltettek az ott végzetteknek*: Frauke Kreuter, Stanley Presser és Roger Tourangeau: „Social Desirability Bias in CATI, IVR, and Web Surveys”, *Public Opinion Quarterly* 72(5), 2008.

127. oldal *a közvélemény-kutatások nem jelezték előre*: Arra vonatkozóan, hogy a hazugság problémát jelenthet Trump támogatótságának előrejelzésekor, ld. Thomas B. Edsall: „How Many People Support Trump But Don’t Want to Admit It?”, *The New York Times*, 2016. május 15, SR 2. Arra vonatkozóan, hogy ez mégsem volt olyan jelentős tényező, ld. Andrew Gelman: „Explanations for That Shocking 2% Shift”, *Statistical Modeling, Causal Inference, and Social Science*, 2016. november 9., <http://andrewgelman.com/2016/11/09/explanations-shocking-2-shift/>.
127. oldal *mondja erről Tourangeau*: Roger Tourangeau-val telefonon készítettem interjút 2015. május 5-én.
128. oldal *miért állítja magáról olyan sok ember, hogy átlagon felüli*: Erről ld. Adam Grant: *Originals: How Non-Conformists Move the World*, New York, Viking, 2016. (Magyarul: *Eredetik. Hogyan mozdítják előre a világot a nonkonformisták?* HVG Könyvek, 2016.) Az eredeti forrás: David Dunning, Chip Heath és Jerry M. Suls: „Flawed Self-Assessment: Implications for Health, Education, and the Workplace”, *Psychological Science in the Public Interest* 5 (2004).
128. oldal *a megkérdezettek egyszerűen szórakoznak*: Anya Kamenetz: „Mischievous Responders’ Confound Research on Teens”, *nprED*, 2014. május 22., <http://www.npr.org/sections/ed/2014/05/22/313166161/mischievous-responders-confound-research-on-teens>. A cikkben tárgyalt eredeti kutatás: Joseph P. Robinson-Cimpian: „Inaccurate Estimation of Disparities Due to Mischievous Responders”, *Educational Researcher* 43, no. 4 (2014).
130. oldal *Többször keresnek például a neten pornót, mint ahányszor megnézik az időjárás-jelentést*: <https://www.google.com/trends/explore?date=all&geo=US&q=porn,weather>.
130. oldal *ismeri be, hogy szokott pornót nézni*: Amanda Hess: „How Many Women Are Not Admitting to Pew That They Watch Porn?”, *Slate*, 2013. október 11., http://www.slate.com/blogs/xx_factor/2013/10/11/pew_online_viewing_study_percentage_of_women_who_watch_online_porn_is_growing.html.
130. oldal *„fasz”, „baszás” vagy „porno”*: Nicholas Diakopoulos: „Sex, Violence, and Autocomplete Algorithms”, *Slate*, 2013.

augusztus 2., http://www.slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html.

132. oldal *3,6-szor nagyobb eséllyel mondják a Google-nak, hogy megbánták*: Becslésem szerint a különböző megfogalmazásokat figyelembe véve Amerikában havonta mintegy 1730 Google-keresésben mondja ki valaki határozottan, hogy már megbánta a gyermekvállalást. Azt viszont csak körülbelül ötvenen írják be, hogy sajnálják, hogy nincs gyerekük. A 45 évnél idősebb amerikaiak közül mintegy 15,9 millióan gyermektelenek, és összesen mintegy 152 millió amerikainak van utóda. Ez azt jelenti, hogy a megfelelő populációból a gyerekesek körülbelül 3,6-szor nagyobb eséllyel adnak hangot utólagos megbánásuknak, mint a gyermektelenek. Természetesen – amint azt a könyvben már említettem, de nem árt még egyszer hangsúlyozni – ilyen „Google-vallomást” csak egy csekély, kis létszámú csoport tesz. Feltehetően azok, akikben elég erős érzelmek dúlnak ahhoz, hogy egy pillanatra elfelejtsék: a Google-től ebben nem várhatnak segítséget.
134. oldal *ahol a legmagasabb a meglegházasság támogatottsága*: A becslések forrása: Nate Silver: „How Opinion on Same-Sex Marriage Is Changing, and What It Means”, FiveThirtyEight, 2013. március 26., http://fivethirtyeight.blogs.nytimes.com/2013/03/26/how-opinion-on-same-sex-marriage-is-changing-and-what-it-means/?_r=0.
134. oldal *A nemi preferenciájukat megadó férfi facebookozók 2,5 százaléka állítja azt, hogy a férfiak érdeklik*: A Facebook-hirdetések adatainak a szerző által elvégzett elemzése. Nem vettem figyelembe azokat a Facebook-felhasználókat, akik azt adják meg preferenciaként, hogy „férfiak és nők”. Az elemzésem azt sejteti, hogy nem elhanyagolható köztük azok százalékos aránya, akik úgy értelmezik, hogy a kérdés nem romantikus érdeklődésre, hanem baráti kapcsolatra vonatkozik.
136. oldal *A férfiak által indított pornókereséseknek [...] mintegy 5 százaléka irányul férfi melegpornóra*: Amint arról már szó esett, a Google Trends nem bontja a kereséseket nemek szerint. A Google AdWords bontja ugyan a megtekintések számát különféle nemi kategóriák szerint, de ezek nem túl precíz adatok.

Hogy meg tudjam becsülni a keresések nemek szerinti megoszlását, először államonként becsülöm meg a keresési adatok segítségével a melegpornó-keresések százalékos arányát. Ezután normalizálom az adatokat a Google AdWords nemek szerinti adataihoz. A genderspecifikus adatokhoz a Pornhub-adatok felhasználásával is hozzájuthatunk, de a Pornhub erősen szelektív minta lehet, hiszen lehetséges, hogy sok melegférfi inkább csak melegpornóra specializálódott oldalakat látogat. A Pornhub-adatok arra utalnak, hogy a férfiak körében alacsonyabb a melegpornó-használat aránya, mint azt a Google-keresések sugallják. Azt viszont megerősítik, hogy nincs erős összefüggés a homoszexualitással szembeni tolerancia, illetve a melegpornó-használat között. Mindezek az adatok, valamint a további jegyzetek megtalálhatók a honlapomon (sethsd.com), a „Sex” címkével ellátott adatmellékletben.

137. oldal *Közülük mintegy 4 százalék vállalja fel nyíltan a Facebookon, hogy meleg*: A szerző saját, a Facebook hirdetési adatain alapuló számítása: 2017. február 8-án, a San Francisco–Oakland–San Jose médiapiacon mintegy 300 középiskolás fiú állította a Facebookon, hogy a férfiak érdeklik, és mintegy 7800 állította azt, hogy a nők.
141. oldal *„Irán nem olyan, mint a maguk országa, nálunk nincsenek homoszexuálisok”*: „‘We Don’t Have Any Gays in Iran,’ Iranian President Tells Ivy League Audience”, Daily Mail.com, 2007. szeptember 25., <http://www.dailymail.co.uk/news/article-483746/We-dont-gays-Iran-Iranian-president-tells-Ivy-League-audience.html>.
141. oldal *„A mi városunkban nincsenek ilyenek”*: Brett Logiurato: „Sochi Mayor Claims There Are No Gay People in the City”, Sports Illustrated, 2014. január 27.
141. oldal *Az internet azonban arról tanúskodik, hogy mind Szocsiban, mind Iránban jelentős az érdeklődés a melegpornó iránt*: A Google AdWords szerint több tízezer keresés irányul a „reñ porno”-ra, azaz melegpornóra. A pornókereséseken belül nagyjából ugyanakkora a melegpornó-keresések aránya Szocsiban, mint az Egyesült Államokban. Iránra vonatkozó adatai nincsenek a Google AdWordsnek, és a Pornhub sem közöl iráni adatokat.

- A PornMD viszont vizsgálta keresési adatait, és arról számolt be, hogy az iráni keresőkifejezések tízes toplistáján öt melegpornóra irányult. Ezek között szerepelt a „daddy love” („apuszerelem”) és a „hotel businessman” („hotelüzletember”). A kutatást ld. Joseph Patrick McCormick: „Survey Reveals Searches for Gay Porn Are Top in Countries Banning Homosexuality”, *PinkNews*, <http://www.pinknews.co.uk/2013/03/13/survey-reveals-searches-for-gay-porn-are-top-in-countries-banning-homosexuality/>.
- A Google Trends szerint az iráni pornókeresések mintegy két százaléka irányul melegpornóra. Az arány kisebb, mint az Egyesült Államokban, de így is széles körű érdeklődésre utal.
144. oldal *a szex tárgykörében*: Stephens-Davidowitz: „Searching for Sex”. A vonatkozó adatok megtalálhatók a honlapomon (sethsd.com), a „Sex” címkéjű adatmellékletben.
144. oldal *nők mintegy 11 százaléka*: Current Contraceptive Status Among Women Aged 15–44: United States, 2011–2013, Járványvédelmi és -Megelőzési Központ, http://www.cdc.gov/nchs/data/databriefs/db173_table.pdf#1.
144. oldal *a 11 százaléknak havonta a 10 százaléka teherbe essen*: David Spiegelhalter: „Sex: What Are the Chances?”, BBC News, 2012. március 15., <http://www.bbc.com/future/story/20120313-sex-in-the-city-or-elsewhere>.
144. oldal *113 szülőképes korú nő közül 1: A terhességek száma évente mintegy 6,6 millió, a 15 és 44 év közötti nők száma pedig 62 millió*.
150. oldal *a másik nem orális kényeztetésében*: Mint már említettem, nem tudom, milyen neműek a keresők. Abból indulok ki, hogy a cunnilingusra, azaz a női nemi szerv szájjal történő ingerlésére rákeresők túlnyomó többsége férfi, a fellációra rákeresők túlnyomó többsége pedig nő. Ugyanis az emberek nagy többsége heteró, így az azonos nemű partner kényeztetésével kapcsolatos információkra talán kisebb az igény.
151. oldal *az öt negatív töltetű szót, amelyek a leggyakrabban fordulnak elő*: A Google AdWords adatainak a szerző által elvégzett elemzése.
152. oldal *meg kell ölni őket*: Evan Soltas és Seth Stephens-Davidowitz: „The Rise of Hate Search”, *The New York Times*, 2015.

december 13., SR 1. Az adatok és a további részletek megtalálhatók a honlapomon (sethsd.com), az „Islamophobia” címkéjű adatmellékletben.

155. oldal *tizenhétszer gyakoribb*: A Google Trends-adatoknak a szerző által elvégzett elemzése.
155. oldal *Martin Luther King-emléknap*: A Google Trends-adatoknak a szerző által elvégzett elemzése.
157. oldal *korrelálnak [...] a fehérek és feketék közötti bérkülönbség tárgyában*: Ashwin Rode és Anand J. Shukla: „Prejudicial Attitudes and Labor Market Outcomes”, stencilezett anyag, 2013.
158. oldal *a szülei*: Seth Stephens-Davidowitz: „Google, Tell Me. Is My Son a Genius?”, *The New York Times*, 2014. január 19., SR 6. A pontos keresőkifejezésekre vonatkozó adatok megtalálhatók a Google AdWords segítségével. Becsléseket a Google Trendszel is találhatunk, ha összehasonlítjuk a „gifted” („tehetséges”) és a „son” („fiúgyermek”) keresőszavak gyakoriságát a „gifted” és a „daughter” („lánygyermek”) szavakéval. Például: <https://www.google.com/trends/explore?date=all&geo=US&q=gifted%20son,gifted%20daughter> and <https://www.google.com/trends/explore?date=all&geo=US&q=overweight%20son,overweight%20daughter>. Az általános szabályszerűség alól, mely szerint az emberek több kérdést tesznek fel a fiuk eszére, illetve a lányuk testére vonatkozóan, egyetlen kivétel van: a „fat son”-ra („kövér fiú”) többen kerestek rá, mint a „fat daughter”-re („kövér lány”). Ez a jelek szerint összefügg a vérfertőző pornókeresések népszerűségével, amelyről már esett szó. A „fat” és a „son” szavakat tartalmazó keresések mintegy 20 százalékaiban szerepel a „porn” szó is.
158. oldal *a lányok arányosan 9 százalékkal nagyobb valószínűséggel kerülnek be valamilyen tehetséggondozó programba*: „Gender Equity in Education: A Data Snapshot” (Nemek szerinti egyenlőség az oktatásban: pillanatfelvétel az adatokról), az Egyesült Államok Oktatásügyi Minisztériuma, Állampolgári Jogok Hivatala, 2012. június, <http://www2.ed.gov/about/offices/list/ocr/docs/gender-equity-in-education.pdf>.
159. oldal *a lányok mintegy 28 százaléka túlsúlyos, míg a fiúknak mintegy 35 százaléka*: Gyermek- és Serdülőegészségügyi Adat-

- forrásközpont, <http://www.childhealthdata.org/browse/survey/results?q=2415&g=455&a=3879&r=1>.
160. oldal *Stormfront-profil*: Stephens-Davidowitz: „The Data of Hate”. A vonatkozó adatok letölthetők a sethsd.com oldalon, a „Stormfront” címkéjű adatmellékletben.
163. oldal *nem nőtt a Stormfront iránti érdeklődés, amikor Donald Trump elnökjelölt lett*: a Stormfront iránti, Google-keresésekben megnyilvánuló érdeklődés 2016 októberében hasonló szinten mozgott, mint 2015 októberében. Ez éles kontrasztban áll az Obama első megválasztásakor kialakult helyzettel. 2008 októberében a Google-keresések tanúsága szerint csaknem 60 százalékkal nőtt a Stormfront iránti érdeklődés az előző év októberéhez képest. Obama megválasztásának másnapján pedig mintegy tízszeresére nőtt a Stormfrontra irányuló Google-keresések száma. Ezzel szemben Trump megválasztásának másnapján mintegy két és félszeresére nőtt a Stormfront-keresések száma. Ez nagyjából megegyezik a George W. Bush 2004-es megválasztását követő napon mért adattal, és nem kizárt, hogy nagyrészt a „politikafüggők” fokozott hírehségéből ered.
165. oldal *az internet [...] politikai szegregáció melegágya*: Matthew Gentzkow és Jesse M. Shapiro: „Ideological Segregation Online and Offline”, *Quarterly Journal of Economics* 126, no. 4 (2011).
168. oldal *Facebook-barátaink*: Eytan Bakshy, Solomon Messing és Lada A. Adamic: „Exposure to Ideologically Diverse News and Opinion on Facebook”, *Science* 348, no. 6239 (2015). A kutatók megállapították, hogy az ideológiai hovatartozását közlő, 9 százaléknyi aktív facebookozónak mintegy 23 százalékban ellentétes pártállású barátai vannak a hovatartozásukat szintén feltüntetők között, míg a Facebookon általuk látott hírányagok 28,5 százaléka tükröz ellentétes ideológiát. Ezek a számok nem vethetők össze közvetlenül a szegregációval kapcsolatos más számadatokkal, mert csak a pártállásukat feltüntető facebookozók kis létszámú mintáján alapulnak. Ők feltehetően sokkal aktívabbak politikailag, és hasonló pártállású, politikailag szintén aktív felhasználókkal létesítenek kapcsolatokat. Ha ez igaz, akkor az összképnek az összes felhasználóra kivetítve ennél is sokkal vegyesebbnek kell lennie.

168. oldal *van egy kulcsfontosságú tényező, amely miatt a Facebookon*: A közösségi média részben azért ilyen meglepően sokszínű, mert külön bónusszal jutalmazza a kiemelkedően népszerű, széles körben megosztott cikkeket, függetlenül azok politikai irányultságától. Ld. Solomon Messing és Sean Westwood: „Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online”, 2014.
168. oldal *több barátjuk van a Facebookon, mint offline*: Ld. Ben Quinn: „Social Network Users Have Twice as Many Friends Online as in Real Life”, *The Guardian*, 2011. május 8. A cikk a Cystic Fibrosis Trust 2011-es tanulmányát ismerteti, amely szerint a közösségi háló átlagos felhasználójának 121 online barátja és 55 hús-vér barátja van. A Pew Research 2014-es tanulmánya szerint az átlagos facebookozónak több mint 300 barátja van. Ld. Aaron Smith: „6 New Facts About Facebook”, 2014. február 3., <http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/>.
168. oldal *lazább kötelékek*: Eytan Bakshy, Itamar Rosenn, Cameron Marlow és Lada Adamic: „The Role of Social Networks in Information Diffusion”, *Proceedings of the 21st International Conference on World Wide Web*, 2012.
170. oldal *„A világvége-jóslatok nem jöttek be”*: „Study: Child Abuse on Decline in U. S.”, Associated Press, 2011. december 12.
170. oldal *tényleg kevesebb gyereket bántalmaztak?*: Ld. Seth Stephens-Davidowitz: „How Googling Unmasks Child Abuse”, *The New York Times*, 2013. július 14., SR 5, és Seth Stephens-Davidowitz: „Unreported Victims of an Economic Downturn”, stencilezett anyag, 2013.
171. oldal *elunták a hosszú várakozást, és feladták a dolgot*: „Stopping Child Abuse: It Begins With You”, *The Arizona Republic*, 2016. március 26.
172. oldal *a terhességmegszakítás nem hivatalos módjainak*: Seth Stephens-Davidowitz: „The Return of the D. I. Y. Abortion”, *The New York Times*, 2016. március 6., SR 2. Az adatok és a további részletek megtalálhatók a honlapomon (sethsd.com), a „Self-Induced Abortion” címkéjű adatmellékletben.
175. oldal *Mindkét kiadvány átlagos példányszáma nagyjából ha-*

- sonló: Alliance for Audited Media, Consumer Magazines, <http://abc3.auditedmedia.com/ecirc/magtitlesearch.asp>.
175. oldal a Facebookon: A szerző számításai a Facebook Ads Manager segítségével, 2016. október 4-én.
176. oldal tíz leglátogatottabb webhely között: „List of Most Popular Websites”, Wikipédia. A böngészők viselkedését nyomon követő Alexa szerint 2016. szeptember 4-én az XVideos volt a legnépszerűbb pornóoldal, egyben az 57. legnépszerűbb weboldal. A SimilarWeb szerint 2016. szeptember 4-én az XVideos volt a legnépszerűbb pornóoldal, egyben a 17. legnépszerűbb weboldal. Az Alexa szerint így néz ki a tízes toplist: Google, YouTube, Facebook, Baidu, Yahoo!, Amazon, Wikipedia, Tencent QQ, Google India, Twitter.
178. oldal 2006. szeptember 5-én kora reggel: A történet forrása: David Kirkpatrick: *The Facebook Effect: The Inside Story of the Company That Is Connecting the World*, New York, Simon & Schuster, 2010. (Magyarul: *A Facebook-jelenség. Egy bennfentes sztorija a cégről, amely összeköti a világot*, Gabo, 2018.)
180. oldal a nagy vállalkozások mindig titkokra épülnek: Peter Thiel és Blake Masters: *Zero to One: Notes on Startups, or How to Build the Future*, New York, The Crown Publishing Group, 2014. (Magyarul: *Nulláról az egyre. Gondolatok startupokról, avagy hogyan teremtsünk jövőt*, Gabo, 2015.)
182. oldal mondja Xavier Amatriain: Xavier Amatriainnal telefonon készítettem interjút 2015. május 5-én.
185. oldal azokat a kérdéseket, amelyek Obama 2014-es évvértékelő beszéde alatt a leginkább foglalkoztatták az amerikaiakat: A Google Trends adatainak a szerző által elvégzett elemzése.
189. oldal ezúttal egy mecsetben: „The President Speaks at the Islamic Society of Baltimore”, YouTube-videó, posztolva 2016. február 3-án, <https://www.youtube.com/watch?v=LRRVdVqAjdW>.
189. oldal elhangzása utáni órákban csökkent a muszlimokkal szembeni gyűlöletet, indulatokat tükröző keresések száma: A Google Trends adatainak a szerző által elvégzett elemzése. A „muszlimokat ölni” keresőkifejezés ritkább volt, mint az előző hét hasonló időszakában. Sőt, az olyan keresések száma is visszaesett, amelyekben együtt szerepelt a „muszlimok” szó, valamint a csoportra aggatott öt leggyakoribb negatív jelző egyike.

5. FEJEZET: ZOOMOLJUNK RÁ!

192. oldal miként befolyásolják gyerekkori élményeink a baseball-csapat-választásunkat: Seth Stephens-Davidowitz: „They Hook You When You’re Young”, *The New York Times*, 2014. április 20., SR 4. A vonatkozó adatok, illetve a kódolás megtalálható a honlapomon (sethsd.com), a „Baseball” címkéjű adatmellékletben.
197. oldal mikor alakulnak ki a legnagyobb valószínűséggel a politikai nézetek: Yair Ghitza és Andrew Gelman: „The Great Society, Reagan’s Revolution, and Generations of Presidential Voting”, kiadatlan kézirat.
199. oldal magyarázza Chetty: Raj Chettyvel telefonon készítettem interjút 2015. július 30-án.
202. oldal megszökjenek „a Nagy Kaszás” elől: Raj Chetty et al.: „The Association Between Income and Life Expectancy in the United States, 2001–2014”, *JAMA* 315, no. 16 (2016).
204. oldal a jelenség mögött részben talán az állhat, hogy az életmód ragályos: Julia Belluz: „Income Inequality Is Chipping Away at Americans’ Life Expectancy”, *vox.com*, 2016. április 11.
204. oldal miért csálnak egyesek adót: Raj Chetty, John Friedman és Emmanuel Saez: „Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings”, *American Economic Review* 103, no. 7 (2013).
207. oldal elhatároztam, hogy letöltöm a Wikipédiát: Seth Stephens-Davidowitz: „The Geography of Fame”, *The New York Times*, 2014. március 23., SR 6. A vonatkozó adatok megtalálhatók a honlapomon (sethsd.com), a „Wikipedia Birth Rate, By County” címkéjű adatmellékletben. A Wikipedia-szócikkek letöltésében, illetve a születési megye kódolásában nyújtott segítségéért köszönetet mondok Noah Stephens-Davidowitznak.
210. oldal az adott megyében van nagyváros: A nagyváros előnyei-re vonatkozóan további bizonyítékokkal szolgál a következő könyv: Ed Glaeser: *Triumph of the City*, New York, Penguin, 2011. (Glaeser a konzulensem volt a posztgraduális képzésen.)
218. oldal számos példa van arra, hogy a való élet utánozza a művészetet: David Levinson, szerk.: *Encyclopedia of Crime and Punishment*, Thousand Oaks, CA: SAGE, 2002.
219. oldal az alanyok egy erőszakos film hatására felfokozott indulatokról számolnak be: Craig Anderson et al.: „The Influence of

- Media Violence on Youth", *Psychological Science in the Public Interest* 4 (2003).
219. oldal azokon a hétvégéken, amikor a mozikba kerül egy-egy erőszakos film: Gordon Dahl és Stefano DellaVigna: „Does Movie Violence Increase Violent Crime?”, *Quarterly Journal of Economics* 124, no. 2 (2009).
221. oldal A Google-kereséseket is ki lehet gyűjteni percenkénti bontásban: Seth Stephens-Davidowitz: „Days of Our Digital Lives”, *The New York Times*, 2015. július 5., SR 4.
223. oldal a bűncselekmények egyik fő mozgatórugója sokszor az alkohol: Anna Richardson és Tracey Budd: „Young Adults, Alcohol, Crime and Disorder”, *Criminal Behaviour and Mental Health* 13, no. 1 (2003); Richard A. Scribner, David P. MacKinnon és James H. Dwyer: The Risk of Assaultive Violence and Alcohol Availability in Los Angeles County”, *American Journal of Public Health* 85, no. 3 (1995); Dennis M. Gorman, Paul W. Speer, Paul J. Gruenewald és Erich W. Labouvie: „Spatial Dynamics of Alcohol Availability, Neighborhood Structure and Violent Crime”, *Journal of Studies on Alcohol* 62, no. 5 (2001); Tony H. Grubestic, William Alex Pridemore, Dominique A. Williams és Loni Philip-Tabb: „Alcohol Outlet Density and Violence: The Role of Risky Retailers and Alcohol-Related Expenditures”, *Alcohol and Alcoholism* 48, no. 5 (2013).
224. oldal hagyja amerikai futballozni mind a négy fiát: „Ed McCaffrey Knew Christian McCaffrey Would Be Good from the Start – ‘The Herd’”, YouTube-videó, posztolva 2015. december 3-án, <https://www.youtube.com/watch?v=boHMmp7DpX0>.
224. oldal az óriási adathalmaz segítségével: A kutatók még sok mindenre rájöttek a kis időintervallumokra lebontott bűnözési adatok elemzése révén. Csak egy példa: az olyan amerikaifoci-meccsek után, amelyeket egy város csapata annak ellenére is elveszít, hogy illett volna megnyernie, gyakoribbak a családon belüli erőszakkal kapcsolatos panaszok. Ld. David Card és Gordon B. Dahl: „Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior”, *Quarterly Journal of Economics* 126, no. 1 (2011).
225. oldal Bill Simmons [...] így írta le: Bill Simmons: „It’s Hard to

- Say Goodbye to David Ortiz”, ESPN.com, június 2., 2009, <http://www.espn.com/espnmag/story?id=4223584>.
226. oldal miként jósolhatjuk meg egy baseballjátékos jövőbeni teljesítményét: A téma kifejtését ld. Nate Silver: *The Signal and the Noise: Why So Many Predictions Fail – But Some Don’t*, New York, Penguin, 2012.
227. oldal a nagydarab sztárfutballisták valóban korán eljutnak a csúcsra: Ryan Campbell: „How Will Prince Fielder Age?”, 2011. október 28., <http://www.fangraphs.com/blogs/how-will-prince-fielder-age/>.
228. oldal Ortiz alteregói: Az adatokat Rob McQuown, a Baseball Prospectus munkatársa volt szíves rendelkezésemre bocsátani.
232. oldal teszi fel a kérdést Kohane: Isaac Kohane-nal telefonon készítettem interjút 2015. június 15-én.
233. James Heywood vállalkozó: James Heywooddal telefonon készítettem interjút 2015. augusztus 17-én.

6. FEJEZET: LABOR AZ EGÉSZ VILÁG

236. oldal 2000. február 27-e: A sztorit tárgyalja többek között Brian Christian: „The A/B Test: Inside the Technology That’s Changing the Rules of Business”, *Wired*, 2012. április 25., http://www.wired.com/2012/04/ff_abtesting/.
238. oldal Ott, ahol a tanárok napidíjat kaptak, felére csökkent az elmaradt órák száma: Esther Duflo, Rema Hanna és Stephen P. Ryan: „Incentives Work: Getting Teachers to Come to School”, *American Economic Review* 102, no. 4 (2012).
238. oldal Bill Gatest annyira lenyűgözték Esther Duflo eredményei: Ian Parker: „The Poverty Lab”, *New Yorker*, 2010. május 17.
240. oldal A Google mérnökei 2011-ben hétezer A/B-tesztet futtattak le: Brian Christian: „The A/B Test”.
240. oldal a kék szín negyvenegy, egymástól alig megkülönböztethető árnyalatát: Douglas Bowman: „Goodbye, Google”, stopdesign, 2009. március 20., <http://stopdesign.com/archive/2009/03/20/goodbye-google.html>.
240. oldal A Facebook ma már napi ezer A/B-tesztet futtat: Eytan Bakshy: „Big Experiments: Big Data’s Friend for Making De-

- cisions”, 2014. április 3., <https://www.facebook.com/notes/facebook-data-science/big-experiments-big-datas-friend-for-making-decisions/10152160441298859/>. A gyógyszeripari tanulmányokkal kapcsolatos információ forrásait ld. „How many clinical trials are started each year?”, Quora-poszt, <https://www.quora.com/How-many-clinical-trials-are-started-each-year>.
241. oldal *Optimizely*: Dan Sirokerrel telefonon készítettem interjút 2015. április 29-én.
242. oldal *nagyjából 60 millió dollár pluszadomány futott be*: Dan Siroker: „How Obama Raised \$60 Millió by Running a Simple Experiment”, Optimizely blog, 2010. november 29., <https://blog.optimizely.com/2010/11/29/how-obama-raised-60-millio-by-running-a-simple-experiment/>.
244. oldal *A The Boston Globe A/B-teszteli a szalagcímeit*: A *The Boston Globe* A/B-tesztjeit és azok eredményeit a szerző rendelkezésére bocsátották. A tesztekéről részletesebb információ található itt: „The Boston Globe: Discovering and Optimizing a Value Proposition for Content”, Marketing Sherpa videóarchívuma, <https://www.marketingsherpa.com/video/boston-globe-optimization-summit2>. A videón Peter Doucette-nek, a *Globe* munkatársának és Pamela Markeynak, a MECLABS munkatársának a beszélgetése hallható.
247. oldal *Benson [...]* így nyilatkozott erről: Clark Bensonnal telefonon készítettem interjút 2015. július 23-án.
247. oldal *hozzájuk csapott egy kockában elhelyezett, jobbra mutató nyilat*: „Enhancing Text Ads on the Google Display Network”, Inside Ad-Sense, 2012. december 3., <https://adsense.googleblog.com/2012/12/enhancing-text-ads-on-google-display.html>.
248. oldal *sok Google-felhasználó is kritikusan viszonyult hozzá*: Ld. pl.: „Large arrows appearing in google ads – please remove”, DoubleClick Publisher Help Forum, https://productforums.google.com/forum/#!topic/dfp/p_TRMqWUF9s.
249. oldal *manapság társadalmunkban elszaporodtak az addiktív viselkedésformák*: Adam Alter: *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*, New York: Penguin, 2017. (Magyarul: *Ellenállhatatlan. Hogyan okoz függőséget a technológia?*, HVG Könyvek, 2018.)
249. oldal *a Google-nak leggyakrabban bevallott függőségek*: A Google Trends-adatoknak a szerző által elvégzett elemzése.
252. oldal *mondta egy előadásában Levitt*: A témáról szó esik egy videóban, amely jelenleg is látható a Harry Walker Speakers Bureau Lökönómia-oldalán: <http://www.harrywalker.com/speakers/authors-of-freakonomics/>.
255. oldal *a Super Bowl alatt sugárzott sör- és üdítőreklámokat vizsgálták*: Wesley R. Hartmann és Daniel Klapper: „Super Bowl Ads”, kiadatlan kézirat, 2014.
256. oldal *miként reagál egy ország a vezetője elvesztésére*: Benjamin F. Jones és Benjamin A. Olken: „Hit or Miss? The Effect of Assassinations on Institutions and War”, *American Economic Journal: Macroeconomics* 1, no. 2 (2009).
257. oldal *pattanásos, alsógatyás kölyök*: Az amellet szóló erős érveket, hogy valószínűleg egy számítógépes szimulációban élünk, ld. Nick Bostrom: „Are We Living in a Computer Simulation?”, *Philosophical Quarterly* 53, no. 211 (2003).
257. oldal *43 amerikai elnök közül*: *Los Angeles Times*, szerkesztőségi cikk: „U. S. Presidential Assassinations and Attempts”, *Los Angeles Times*, 2012. január 22., <http://timelines.latimes.com/us-presidential-assassinations-and-attempts/>.
257. oldal *Vessük össze például John F. Kennedy és Ronald Reagan esetét*: Benjamin F. Jones és Benjamin A. Olken: „Do Assassins Really Change History?”, *The New York Times*, 2015. április 12, SR 12.
258. oldal *csak Kadirov halt meg*: A támadásról elérhető egy felkavaró videó: „Parade surprise (Chechnya 2004)”, YouTube-videó, posztolva 2009. március 31-én, <https://www.youtube.com/watch?v=fHWhs5QkfuY>.
258. oldal *Hitler ugyanis módosította a programját*: Ezt az esetet is tárgyalja Jones és Olken a „Do Assassins Really Change History?” című cikkben.
260. oldal *a lottónyeremény rövidtávon nem*: A megállapítást ld. John Tierney: „How to Win the Lottery (Happily)”, *The New York Times*, 2015. május 27., D5. Tierney cikke a következő tanulmányokat tárgyalja: Bénédicte Apouey és Andrew E. Clark: „Winning Big but Feeling No Better? The Effect of Lottery Prizes on Physical and Mental Health”, *Health Economics* 24, no.

- 5 (2015); Jonathan Gardner és Andrew J. Oswald: „Money and Mental Wellbeing: A Longitudinal Study of Medium-Sized Lottery Wins”, *Journal of Health Economics* 26, no. 1 (2007); és Anna Hedenus: „At the End of the Rainbow: Post-Winning Life Among Swedish Lottery Winners”, kiadatlan kézirat, 2011. Tierney arra is rámutat cikkében, hogy a híres 1978-as tanulmány, amely szerint a lottónyeremény nem tesz boldoggá, (Philip Brickman, Dan Coates és Ronnie Janoff-Bulman: „Lottery Winners and Accident Victims: Is Happiness Relative?”, *Journal of Personality and Social Psychology* 36, no. 8 (1978) nagyon kis létszámú mintával dolgozott.
260. oldal *a szomszéd lottónyereményének*: Ld. Peter Kuhn, Peter Koorernan, Adriaan Soeteven és Arie Kapteyn: „The Effects of Lottery Prizes on Winners and Their Neighbors: Evidence from the Dutch Postcode Lottery”, *American Economic Review* 101, no. 5 (2011); illetve Sumit Agarwal, Vyacheslav Mikhed és Barry Scholnick: „Does Inequality Cause Financial Distress? Evidence from Lottery Winners and Neighboring Bankruptcies”, munkaanyag, 2016.
260. oldal *a lottónyertesek szomszédai*: Agarwal, Mikhed és Scholnick: „Does Inequality Cause Financial Distress?”
261. oldal *az orvosokat motiválhatják az anyagiak*: Jeffrey Clemens és Joshua D. Gottlieb: „Do Physicians’ Financial Incentives Affect Medical Treatment and Patient Health?”, *American Economic Review* 104, no. 4 (2014). Fontos, hogy ezek a kutatási eredmények nem azt bizonyítják, hogy az orvosok gonoszak. Igazság szerint aggasztóbb lenne, ha az olyan eljárások, amelyeket az orvosok csak akkor rendelnek el, ha többet fizetnek értük, valóban életet mentenének. Ha így lenne, az azt jelentené, hogy az orvosokat jól meg kell fizetni ahhoz, hogy elrendeljék az életmentő kezeléseket. Clemens és Gottlieb eredményei azonban nem ezt sugallják. Hanem azt, hogy az életmentő kezeléseket a pénztől függetlenül, mindenképpen elrendelik, az olyan eljárásokat viszont, amelyek nem segítenek túl sokat, csak akkor, ha jól fizetnek értük. Megfogalmazhatnánk ugyanezt így is: az orvosokat nemigen érdeklik az anyagiak, ha életmentő beavatkozásról van szó, a kevésbé fontos kezeléseknél azonban nagyon is motiválja őket a pénz.
261. oldal *150 millió dollárt érő*: Robert D. McFadden és Eben Shapiro: „Finally, a Face to Fit Stuyvesant: A High School of High Achievers Gets a High-Priced Home”, *The New York Times*, 1992. szeptember 8.
262. oldal *A diákok ötvenötfféle*: A kurzusok listája hozzáférhető a Stuy honlapján: <http://stuy.enschool.org/index.jsp>.
262. oldal *az itt végzettként egy negyedét veszítik fel*: Anna Bahr: „When the College Admissions Battle Starts at Age 3”, *The New York Times*, 2014. július 29., <http://www.nytimes.com/2014/07/30/upshot/when-the-college-admissions-battle-starts-at-age-3.html>.
262. oldal *A Stuyvesant Középiskolában tanult*: Sewell Chan: „The Obama Team’s New York Ties”, *The New York Times*, 2008. november 25.; Evan T. R. Rosenman: „Class of 1984: Lisa Randall”, *Harvard Crimson*, 2009. június 2.; „Gary Shteyngart on Stuyvesant High School: My New York”, YouTube-videó, posztolva 2010. augusztus 4-én, https://www.youtube.com/watch?v=NQ_phGkC-Tk; Candace Amos: „30 Stars Who Attended NYC Public Schools”, *New York Daily News*, május 29., 2015.
262. oldal *Tanévzáró beszédet pedig mondott már*: Carl Campanile: „Kids Stuy High Over Bubba: He’ll Address Ground Zero School’s Graduation”, *New York Post*, 2002. március 22.; ENSZ-sajtóközlemény: „Stuyvesant High School’s ‘Multicultural Tapestry’ Eloquent Response to Hatred, Says Secretary-General in Graduation Address”, 2004. június 23.; „Conan O’Brien’s Speech at Stuyvesant’s Class of 2006 Graduation in Lincoln Center”, YouTube-videó, posztolva 2012. május 6-án, <https://www.youtube.com/watch?v=zAMkUE9Oxnc>.
262. oldal *a Stuy az első helyen végzett*: Ld. <https://k12.niche.com/rankings/public-high-schools/best-overall/>.
263. oldal *alig 5 százaléka kerül be*: Pamela Wheaton: „8th-Graders Get High School Admissions Results”, *Insideschools*, 2016. március 4., <http://insideschools.org/blog/item/1001064-8th-graders-get-high-school-admissions-results>.
267. oldal *ha valaki durvább körülmények között raboskodik*: M. Keith Chen és Jesse M. Shapiro: „Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach”, *American Law and Economics Review* 9, no. 1 (2007).

267. oldal *Mekkora hatása volt a Stuyvesant Középiskolának?*: Atilla Abdulkadiroglu, Joshua Angrist és Parag Pathak: „The Elite Illusion: Achievement Effects at Boston and New York Exam Schools”, *Econometrica* 82, no. 1 (2014). Ugyanezt az eredményt, azaz egy nagy nullát hozott ki a következő tanulmány is: Will Dobbie és Roland G. Fryer Jr.: „The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools”, *American Economic Journal: Applied Economics* 6, no. 3 (2014).
269. oldal *Az átlagos Harvard-diplomás*: Ld. <http://www.payscale.com/college-salary-report/bachelors>.
269. oldal *a hasonlóan nagy presztízssű egyetemre felvett, de másikat választó diák nagyjából ugyanoda jut el, mint az, aki az elit egyetem hallgatója lesz*: Stacy Berg Dale és Alan B. Krueger: „Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables”, *Quarterly Journal of Economics* 117, no. 4 (2002).
271. oldal *Warren Buffett*: Alice Schroeder: *The Snowball: Warren Buffett and the Business of Life*, New York, Bantam, 2008.

7. FEJEZET:

NAGY ADAT, NAGY GOND? AMI NEM MEGY

280. oldal *azt állította, hogy az emberek Twitter-bejegyzései alapján előre lehet jelezni, merre mozdul el*: Johan Bollen, Huina Mao és Xiaojun Zeng: „Twitter Mood Predicts the Stock Market”, *Journal of Computational Science* 2, no. 1 (2011).
280. oldal *A Twitter-alapú hedge fundot pedig egy hónappal az alapítása után megszüntették*: James Mackintosh: „Hedge Fund That Traded Based on Social Media Signals Didn't Work Out”, *Financial Times*, 2012. május 25.
282. oldal *egyszer sem tudták reprodukálni a korrelációt*: Christopher F. Chabris et al.: „Most Reported Genetic Associations with General Intelligence Are Probably False Positives”, *Psychological Science* (2012).
285. oldal *Zoë Chance: A történetet ld. „How to Make a Behavior Addictive: Zoë Chance at TEDx Mill River”, YouTube-videó, posztolva 2013. május 14-én*, <https://www.youtube.com/watch?v=AHfiKav9fcQ>.

v=AHfiKav9fcQ. Egyes részletek, így a lépcsőszámláló színe is, az interjúk során derültek ki. Zoë Chance-szal 2015. április 20-án telefonon, 2016. július 11-én és 2016. szeptember 8-án pedig e-mailen készítettem interjút.

286. oldal *a számok bódítóak tudnak lenni*: Az alfejezet forrása: Alex Peysakhovich és Seth Stephens-Davidowitz: *How Not to Drown in Numbers*, *The New York Times*, május 3., 2015, SR 6.
286. oldal *egyes pedagógusok egyenesen csálnak a tesztekkel*: Brian A. Jacob és Steven D. Levitt: „Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating”, *Quarterly Journal of Economics* 118, no. 3 (2003).
289. oldal *véli Thomas Kane*: Thomas Kane-nel telefonon készítettem interjút 2015. április 22-én.
289. oldal *„Mindegyik mérési módszer hozzászól valami értékeset a kutatáshoz”*: Bill és Melinda Gates Alapítvány: „Ensuring Fair and Reliable Measures of Effective Teaching”, http://k12education.gatesfoundation.org/wp-content/uploads/2015/05/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

8. FEJEZET:

NAGY ADAT, NAGY GOND? AMIT NEM KELLENE

291. oldal *A közelmúltban három közgazdász*: Oded Netzer, Alain Lemaire és Michal Herzenstein: „When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications”, 2016.
291. oldal *a hitelek mintegy 13 százaléka dől be*: Peter Renton: „Another Analysis of Default Rates at Lending Club and Prosper”, 2012. október 25., <http://www.lendacademy.com/lending-club-prosper-default-rates/>.
295. oldal *a Facebook-lájkok gyakran korrelálnak*: Michal Kosinski, David Stillwell és Thore Graepel: „Private Traits and Attributes Are Predictable from Digital Records of Human Behavior”, *PNAS* 110, no. 15 (2013).
299. oldal *a cégek mennyire ki vannak szolgáltatva a Yelp-értékelések kényének-kedvének*: Michael Luca: „Reviews, Reputation, and Revenue: The Case of Yelp”, kiadatlan kézirat, 2011.
301. oldal *öngyilkossággal kapcsolatos Google-keresések*: Christine Ma-Kellams, Flora Or, Ji Hyun Baek és Ichiro Kawachi: „Ret-

hinking Suicide Surveillance: Google Search Data and Self-Reported Suicidality Differentially Estimate Completed Suicide Risk", *Clinical Psychological Science* 4, no. 3 (2016).

302. oldal 3,5 millió Google-keresés: Az alkalmazott módszert tárgyalom a honlapomon, a házilag végrehajtott abortusszal kapcsolatos jegyzetekben. Az „öngyilkosság” Google-kategóriára irányuló kereséseket vettem össze a „hogyan kössünk nyakkendőt” keresésekkel. 2015-ben 6,6 milliószor fordult elő a „hogyan kössünk nyakkendőt” Google-keresőkifejezés. Az „öngyilkosság” kategóriába tartozó keresések száma ennek 6,5-szerese volt. $6,5\text{-szer } 6,6\text{, osztva } 12\text{-vel} = 3,5$.

303. oldal 12 olyan, muszlimok ellen elkövetett gyilkosság történt, amelyet utóbb gyűlöletbűncselekménynek minősítettek: Bridge Initiative Team: „When Islamophobia Turns Violent: The 2016 U. S. Presidential Election”, 2016. május 2., hozzáférhető: <http://bridge.georgetown.edu/when-islamophobia-turns-violent-the-2016-u-s-presidential-elections/>.

ZÁRSZÓ

307. oldal Mi motiválta Popper keresztes hadjáratát? Karl Popper: *Conjectures and Refutations* (Sejtések és cáfolatok), London, Routledge & Kegan Paul, 1963.

310. oldal a város térképén bejelölte az összes koleraesetet: Simon Rogers: „John Snow’s Data Journalism: The Cholera Map That Changed the World”, *The Guardian*, 2013. március 15.

311. oldal Benjamin F. Jones: Benjamin Jonesszal telefonon készítettem interjút 2015. június 1-én. A projekt leírását ld. még: Aaron Chatterji és Benjamin Jones: „Harnessing Technology to Improve K-12 Education”, Hamilton Project, vitaanyag, 2012.

319. oldal nemigen jellemző, hogy az emberek átrágják magukat a közgazdászok terjedelmes tanulmányain: Jordan Ellenberg: „The Summer’s Most Unread Book Is...”, *Wall Street Journal*, 2014. július 3.



Tartalom

Előszó	7
Bevezetés	11
I. RÉSZ	
NAGY ADAT, KIS ADAT	
1. fejezet	
Ösztönös „félreérzéseink”	39
II. RÉSZ	
A BIG DATA HATALMA	
2. fejezet	
Igaza volt-e Freudnak?	61
3. fejezet	
Adatok újragondolva	72
4. fejezet	
Digitális igazságszérum	125
5. fejezet	
Zoomoljunk rá!	191
6. fejezet	
Labor az egész világ	236

III. RÉSZ

BIG DATA: VIGYÁZAT, TÖRÉKENY!

7. fejezet

Nagy adat, nagy gond? Ami nem megy 275

8. fejezet

Nagy adat, nagy gond? Amit nem kellene 291

Zárszó: hányan olvasnak végig egy könyvet? 306

Köszönetnyilvánítás 321

Jegyzetek 325

Nő-e az erőszakos bűncselekmények száma a média hatására?

Mennyit szexelnek ténylegesen az emberek?

Kijátszhatjuk-e a tőzsdepiacot?

Hány rasszista élhet valójában az egyes országokban?

Hányan olvassák el a könyvet, amit megvettek?

Honnan tudjuk, hogy az vagy, amire klikkelsz?

Seth Stephens-Davidowitz, a Harvardon végzett közgazdász, a Google volt adattudósa, a *The New York Times* rovatvezetője szerint az, amit az emberekről gondoltunk, nagyrészt totális tévedés. Hogy miért? Azért, mert az emberek hazudnak. Hazudnak a barátaiknak, a szeretőiknek, az orvosuknak, a közvélemény-kutatóknak – és önmaguknak is.

Csakhogy... Az internet korában már nem kell arra hagyatkoznunk, amit az emberek magukról mondanak. A keresőmotorok, a közösségi oldalak, a randi- és a pornóoldalak digitális aranybányák a Big Data kutatóinak. Valós képet adnak arról, mit gondolnak, mit akarnak, mit tesznek valójában az emberek. Ezekből az adatokból megtudjuk, milyenek is vagyunk ténylegesen – ami lehet vicces, de akár sokkoló is, ám mindenképpen elgondolkodtató. Mert a Big Datától szinte mindent megtudhatunk az emberi természetről – feltéve, ha azt kérdezzük tőle, amit kell.

Csatlakozzon közösségünkhöz,
vegyen részt nyereményjátékainkban!

www.facebook.com/athenaeumkiado



9 789632 938684

www.athenaeum.hu 3999 Ft

seth stephens-davidowitz

MINDENKI HAZUDIK



AZ VAGY, AMIRE KLIKKELSZ